THE PROCEEDINGS OF
THE PHYSICAL SOCIETY

Section B

Vol. 64, Part 3 1 March 1951 No. 375 B

The Cohesion of Alloys: I

Intermetallic Systems formed by Copper, Silver and Gold, and Deviations from Vegard's Law

By M. A. JASWON*, W. G. HENRY and G. V. RAYNOR
Metallurgy Department, University of Birmingham

Communicated by H. Jones; MS. received 18th May 1950,
and in final form 11th October 1950

ABSTRACT. In order to extend the theory of the cohesion of metals to alloy structures a new treatment is considered. The excess energy of the solid solution, as compared with that of a mixture of the pure components, is estimated by considering that the introduction of a solute atom causes a major disturbance extending only to its nearest neighbours, and a less serious disturbance in the surrounding matrix. For a face-centred cubic structure, therefore, attention is directed to the group of thirteen atoms formed by the solute atom and its twelve symmetrically situated solvent atoms. The size of the atomic polyhedra forming this group is different from the sizes of the polyhedra in the solvent matrix, and the equilibrium size of the polyhedra within the group is calculated from the known lattice energies of the components, the atomic radii of the components, and the strain energy introduced into the matrix, which is a measure of the long-range minor disturbance. For this purpose it is assumed that the variation with volume of the energies of the components in the alloy is the same as in the pure metals. Thus, the excess energy w associated with the group of thirteen atoms is $w=\{12\epsilon_1(r)+\epsilon_2(r)\}-\{12\epsilon_1(r_1)+\epsilon_2(r_2)\}$, where r_1 and r_2 are the radii of the solvent and solute atoms, and ϵ_1 and ϵ_2 are the cohesive energies of the solvent and solute metals. w may be plotted against r, and the attendant strain energy in the matrix may be similarly plotted. The sum of these energies is minimized to give the equilibrium size of the atomic polyhedra within a group of thirteen atoms. Using these conceptions, the deviations from Vegard's law to be expected in the solid solutions formed by copper, silver and gold taken in pairs are evaluated. The sign of the deviation is correctly predicted in all cases; quantitative agreement between the magnitudes of the expected and observed deviations is obtained for the copper–silver alloys, which provide the most favourable case for treatment. Application of the theory to problems of solid solution formation is briefly discussed. In its present form the theory applies only to dilute solid solutions.

---

§ 1. INTRODUCTION

A considerable amount of work has been carried out in connection with the problem of the cohesion of metals, chiefly those belonging to Groups IA and IB of the Periodic Table, but until recently the cohesion of alloys has received comparatively little attention.

One of the most successful approaches to the problem for pure metals was made by Wigner and Seitz (see Mott and Jones 1936), who, in effect, attribute the whole cohesive energy to the interaction of the ions of the metal with the valency electrons. The whole volume of the metal is considered to be divided geometrically into polyhedra of equal size, which are formed by bisecting the lines joining each atom to its nearest neighbours. The resulting polyhedra have forms which depend upon the crystal structure, and each contains one ionic core, the potential field surrounding which is that of the singly charged positive ion. It is assumed that there is, at any given moment, only one electron in each polyhedron, and that correlation between the electrons may, to a first approximation, be neglected. It is further assumed that the electron clouds of the ions do not overlap, so that the mutual ionic repulsion may be ignored. Since the potential field inside the metal is periodic with the period of the lattice, the Schrödinger equation for the free electron in the potential field of the ion is solved subject to the boundary condition that the resulting wave functions must join up smoothly with each other at the walls of the polyhedra. From the Wigner–Seitz wave function, the energy of the lowest state of the valency electron is obtained as a function of the volume of the polyhedron, and may be represented in the form of curve I of Figure 1. It is found that certain other effects, such as the interaction energy of an electron with other electrons in its own cell, the extra energy due to electrons with parallel spins keeping apart, and the extra energy due to interaction of electrons with anti-parallel spins, approximately cancel out.

---

* Now at the Department of Mathematics, Imperial College, London.

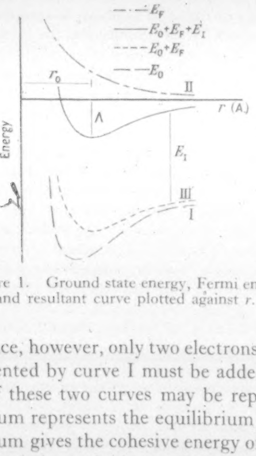


Figure 1. Ground state energy, Fermi energy, and resultant curve plotted against r.

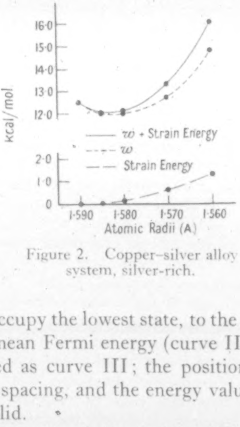


Figure 2. Copper–silver alloy system, silver-rich.

Since, however, only two electrons may occupy the lowest state, to the energies represented by curve I must be added the mean Fermi energy (curve II). The sum of these two curves may be represented as curve III; the position of the minimum represents the equilibrium lattice spacing, and the energy value of the minimum gives the cohesive energy of the solid.

Calculations of this type have been made for the alkali metals by Wigner and Seitz (Wigner and Seitz 1933, 1934, Wigner 1934, Seitz 1935) and for copper and silver by Fuchs (1935) who introduced the additional consideration of the ion–ion repulsion, which is important for the metals of Group IB of the Periodic Table. Agreement between theory and experiment is good for the alkali metals, which constitute favourable cases for this treatment, but less good for copper and silver. In these cases reasonable values are obtained for the lattice spacing and the compressibility, but the calculated binding energy is considerably less than the observed value. The Wigner–Seitz method allows calculation of cohesive energies, therefore, but, owing to the assumptions involved, is not yet quantitatively reliable, It is not to be expected, therefore, that application of the Wigner–Seitz approach to the problem of alloy cohesion would give useful results at present, especially in view of the fact that, when a solute atom replaces a solvent atom, the potential field is no longer perfectly periodic. The disturbance of the potential field by a foreign atom has not been examined in any detail. The only detailed calculations relating to alloys were, until recently, those carried out by Jones (1937) in connection with the stability and composition of electron compounds; in these cases the size relationships between the component metals are favourable, and the effect of alloying is considered to be, essentially, the raising of the Fermi energy by increasing the number of electrons per atom. Arafa (1949), however, has recently applied the general methods of the Wigner–Seitz theory, making allowance for the imperfectly periodic potential in assessing the Fermi energy, to the calculation of heats of solution in the copper–silver system; reasonable results have been obtained. The orthodox wave-mechanical approach, such as that used by Arafa, requires a knowledge of the Hartree ion-core fields of the atoms concerned, and is therefore of somewhat limited application.

A treatment of the problem of deviations from Vegard's law, a problem closely allied to cohesion, has been given by Pines (1940) on the basis of classical elasticity theory. He has determined the change in linear dimensions caused by introducing an elastic sphere having the volume of a solute atom, into a hole which has the volume of a solvent atom in an elastic medium of finite dimensions. Reasonable agreement was obtained between the observed and calculated deviations for a number of systems, including the copper–gold and silver–gold alloys, for which the observed deviations are respectively positive and negative. For the copper–silver alloy system, however, negative deviations are predicted, in contrast to the positive effect observed. Certain assumptions necessary to the treatment appear to limit its accuracy. It is necessary, for example, to define the boundary of a solute atom in solution; in view of the possibly severe disturbance of charge distribution round the solute atom, all traces of the original atomic sizes are lost, so that it becomes difficult to define the boundary as the juncture of the matrix and the introduced elastic sphere. This replacing sphere, representing a single atom, is also considered to be homogeneous, and to have the same elastic constants as the bulk material. Thus, the study of the deviations from Vegard's law on the basis of classical elasticity not only involves a number of assumptions, but gives results which are not entirely consistent with observation.

In connection with a theoretical study of the systems copper–silver, copper–gold, and silver–gold, with particular reference to the lattice spacing relationships, an alternative, semi-empirical approach to the problem of alloy cohesion has been sought. The present paper describes the general principles involved in the method, and the particular application to the problem of deviations from Vegard's law. Certain implications arise from the use of the method which to some extent justify the semi-empirical nature of the approach, as compared with orthodox approaches of more limited application.

§ 2. GENERAL PRINCIPLES

In the following sections, the cohesive energy of the metal is defined as the sum of the sublimation energy and the ionization energy, and represents the total energy per mole of the solid:

$$\epsilon = \Lambda + E_1 \qquad \ldots\ldots(1)$$

where ϵ is the cohesive energy, Λ is the sublimation energy and E_1 the ionization energy.

The energy of a mixture of components 1 and 2, in which the concentration of metal 2 is C may be written

$$\epsilon_p = (1-C)\epsilon_1 + C\epsilon_2 \qquad \ldots\ldots(2)$$

where ϵ_1 and ϵ_2 are the cohesive energies of the pure components. In the case of an alloy of the same composition, the cohesive energy is:

$$\epsilon_0 = (1-C)\epsilon_1 + C\epsilon_2 + C\Delta_{12} \qquad \ldots\ldots(3)$$

where Δ_{12} represents the increase in energy associated with the replacement of a solvent atom by a solute atom in forming a solid solution, over and above that of the mixture of equal composition. The problem therefore becomes that of evaluating the term Δ_{12}.

When a solute atom is transferred from metal 2 to the lattice of metal 1, the disturbance in charge distribution caused by the presence of the solute ion will probably fall off relatively rapidly, affecting its nearest neighbours appreciably, and the surrounding volume to a much smaller degree. This suggests that the disturbance may be analysed into two separate effects as follows: (i) a major local disturbance, of electronic origin, affecting the nearest neighbours to the introduced solute atom, (ii) a minor disturbance, due to longer range electronic effects, in the volume lying outside that occupied by the nearest neighbours.

The conception that the major electronic disturbance due to the solute ion extends only to the nearest neighbours leads to the assumption that the solute atom and its nearest neighbours may be considered together as a group, forming a local 'knot' in the matrix of the solvent. Since this group will, in general, have a different volume from that of the same number of solvent atoms, the minor, long-range disturbance, will appear as strain energy in the matrix, and may be evaluated by the methods of classical elasticity as a function of the size of the group. The energy associated with the local electronic disturbance (i.e. the energy of the group itself) may be assessed approximately, as a function of the size of the group, by a method to be described. The sum of these two energies has a minimum which enables the equilibrium size of the group to be found, and hence the quantity Δ_{12}, which measures the energy change consequent on the introduction of the solute atom, is determinable.

In the method employed to evaluate Δ_{12}, it is necessary to make the assumption that, if atomic polyhedra are described about each of the atoms comprising a group (i.e. one solute atom together with its nearest neighbours) such that they each enclose an ionic core and one valency electron, these polyhedra will be of equal size. This assumption may be regarded as a purely mathematical device to be justified by the results which are obtainable. There is, however, experimental evidence to support the assumption, as may be appreciated from consideration of the observed deviations from Vegard's law in a system such as that formed by the copper–gold alloys. For these alloys a large solute atom replaces a small solvent atom, and yet a positive deviation from Vegard's law is observed. If the solute atom were to retain its characteristic size in the solvent lattice, Vegard's law would be obeyed and, in order to obtain a positive deviation, we must consider either (i) that the solute atom expands to a size greater than that which it has on its own lattice, which is improbable, since on the basis of a simple elastic model compression rather than expansion would be expected, or (ii) that the solute atom is, in fact, slightly compressed, and that the surrounding twelve solvent atoms occupy a slightly larger volume than that characteristic of the bulk material to give a positive resultant deviation.

The second alternative is equivalent to considering that all the atoms of the group tend towards the same size, which is not inconsistent with the initial assumption.

It should be noted that, by considering the group of thirteen atoms, the difficulty of defining the boundary of the solute atom is avoided. All thirteen atoms are assumed to have the same initially unknown size, and this approximation is felt to be justified in the present treatment. There is evidence from the work of Arafa that there is an accumulation of charge in the neighbourhood of a solute atom, a factor which is not explicitly included in the present treatment; this factor is not, however, entirely ignored, since the volumes of the polyhedra which comprise the group of one solute atom and its nearest neighbours are different both from the volume of a polyhedron in the solvent metal and from that of a polyhedron in the solute metal.

The quantity Δ in equation (3) may thus be assessed, provided the solute concentration is small, from a knowledge of the change in the electronic potential energy and the additional elastic energy necessary to replace a solvent atom. If the concentration is not small, then there is a probability that other solute atoms will find themselves included in the group of twelve atoms surrounding a given solute atom, and the present treatment would be inapplicable.

§ 3. THE EQUILIBRIUM SIZE OF POLYHEDRA IN A GROUP

The cohesive energies ϵ_1 and ϵ_2 are well known, in the neighbourhood of their minima, from experimental data. Thus, for instance, Fürth (1944) has deduced a semi-empirical expression for the variation of the energy of a crystal with volume, in which coefficients, characteristic of a given metal, are determined from experimental data involving compressibility, atomic volume, sublimation energy, and temperature and pressure relationships. The use of these curves for the assessment of the energy values involved in alloys appears to be justified, since they are only utilized in the neighbourhood of their minima, and likely to lead to results at least as accurate as any that could be deduced from purely theoretical ground-state curves.

Thus, according to Fürth, the potential energy of the lattice, ϕ, is given by the expression:

$$-\phi = \Lambda(r_0) = qNu_{r_0} \qquad \ldots\ldots(4)$$

where $\Lambda(r_0)$ and u_{r_0} are respectively the sublimation energy per mol of the substance at absolute zero, and the dissociation energy of an isolated pair of atoms, when the atoms are the equilibrium distance apart; q is a function of the co-ordination number. Since it is assumed that each atomic polyhedron in the metal is equivalent, with no interaction between polyhedra, it is reasonable to assume that

$$\Lambda(r)/N = qu_r \qquad \ldots\ldots(5)$$

is the energy to be associated with each polyhedron as a function of the atomic radius r, which is defined by the relation $4\pi r^3/3 = \Omega$, the atomic volume. The form of u_r, according to Fürth, is:

$$u_r = u_{r_0}\frac{nm}{n-m}\left\{-\frac{1}{m}\left(\frac{r_0}{r}\right)^m + \frac{1}{n}\left(\frac{r_0}{r}\right)^n\right\}, \qquad \ldots\ldots(6)$$

where m and n are characteristic of a given metal, and are determined from experimental data. The curve of u_r against r has a minimum at $r=r_0$ and m is always less than n; the minimum is therefore not symmetrical.

The change in electronic potential energy when one solute atom replaces a solvent atom may now be assessed, assuming that each polyhedron in the group is governed by the relation characteristic of the metal as given by Fürth.

From equation (6), the energy to be associated with an electron within an atomic polyhedron is

$$\epsilon(r)=\frac{\Lambda(r_0)}{N}\,\frac{nm}{n-m}\left\{-\frac{1}{m}\left(\frac{r_0}{r}\right)^m+\frac{1}{n}\left(\frac{r_0}{r}\right)^n\right\}. \qquad \ldots\ldots(7)$$

When $r=r_0$, $\epsilon(r_0)=\Lambda(r_0)/N=qu_{r_0}$.

The energy of a group of one solute atom with twelve solvent atoms equidistantly surrounding it may then, according to the assumptions discussed above, be written

$$\epsilon_0(r)=12\epsilon_1(r)+\epsilon_2(r), \qquad \ldots\ldots(8)$$

where ϵ_1 and ϵ_2 refer respectively to the solvent and solute metals. The excess energy w associated with the solid solution is then

$$w=\{12\epsilon_1(r)+\epsilon_2(r)\}-\{12\epsilon_1(r_1)+\epsilon_2(r_2)\}, \qquad \ldots\ldots(9)$$

where r_1 and r_2 are the radii of the solvent and solute atoms.

The group of thirteen atoms may be considered to possess an effective radius R_1 given by the relation:

$$4\pi R_1{}^3/3=13(4\pi r^3/3), \qquad \ldots\ldots(10)$$

whereas a group of thirteen solvent atoms would have an effective radius r_0, given by $4\pi r_0{}^3/3=13(4\pi r_1{}^3/3)$.

In order to assess the minor disturbance in terms of a strain energy factor, the process of solid solution formation may be considered as the introduction into a hole in an elastic medium of radius r_0 of an elastic sphere of some radius R_0, which in equilibrium assumes a radius R_1. Following Frenkel (1946) or Mott and Nabarro (1940), the strain energy in the medium is then:

$$8\pi r_0(R_1-r_0)^2G, \qquad \ldots\ldots(11)$$

where G is the modulus of rigidity. This strain energy is a positive addition to the total energy of the solid solution for both positive and negative values of (R_1-r_0) and is equal to zero at $R_1=r_0$.

The equilibrium size of the atomic polyhedra within the group may be obtained by summation of equations (9) and (11), and the value of w at this value of r may be calculated.

The mean atomic radius of the resulting solid solution may be defined as

$$\rho=(1-13C)r_1+13Cr, \qquad \ldots\ldots(12)$$

where r is the radius of the individual polyhedra within the group of thirteen atoms. If the material consisted of a phase mixture of the components, or alternatively of a solid solution with an additive relationship between the atomic radii of the components, the mean atomic radius would be

$$\rho_1=(1-C)r_1+Cr_2. \qquad \ldots\ldots(13)$$

Equations (12) and (13) are both linear functions of C for dilute solution, so that the deviation from Vegard's law is also a linear function of C. This may be written:

$$\Delta\rho=\rho-\rho_1=Cr_1(13K-K'), \qquad \ldots\ldots(14)$$

where $K=(r-r_1)/r_1$ and $K'=(r_2-r_1)/r_1$. The deviation from Vegard's law is therefore calculable; the rate of change of $\Delta\rho$ with respect to C is compared with experiment in the next section.

§ 4. COMPARISON WITH EXPERIMENT

From the preceding section it will be seen that the deviations from Vegard's law in solid solutions may be assessed from a knowledge of the lattice energies of the solvent and solute metals by considering the equilibrium size of a group of thirteen atoms composed of one solute atom and its twelve nearest neighbours, for the face-centred cubic structure. From (9), w may be plotted against r, assuming that the variations of ϵ_1 and ϵ_2 with r are the same as the variations in the case of pure metals; from (11), the strain energy factor may be plotted against r, and the minimum of the summation curve determines the equilibrium size of the polyhedra in the group. K and K' in equation (14) are thus evaluated, and the deviation from Vegard's law may be assessed.

In this section, numerical data are presented for comparison with experiment. Table 1 gives the constants of the metals used in the calculations.

Table 1

	Λ (kcal/mol.)	E_1 (kcal/mol.)	ϵ (kcal/mol.)	Atomic radius (A.)	m	n	G (dynes/cm^2).
Copper	81·2	177·4	258·6	1·4099	4·0	7	$4{\cdot}55\times10^{11}$
Silver	68·0	174·0	242·0	1·5933	4·5	7	$2{\cdot}87\times10^{11}$
Gold	92·0	212·0	304·0	1·5905	5·5	8	$2{\cdot}77\times10^{11}$

The relevant data for the alloy systems are collected in Table 2. Figures 2 to 7 show graphically the variation of w and strain energy with r for the respective solid solutions; the minima in the summation curve give the radius values in column 6 of Table 2. Table 2 also contains values of Δ_{12} for completeness.

The deviations from Vegard's law which are found experimentally for the systems under consideration are shown in Figures 8, 9 and 10. It has already been shown that the mean atomic radius for a solid solution is:

$$\rho=(1-13C)r_1+13Cr, \qquad \ldots\ldots(12)$$

where r is the radius of the atom in a group. Putting

$$r=(12r_1+r_2)/13 \qquad \ldots\ldots(15)$$

then

$$\rho=(1-13C)r_1+13C(12r_1+r_2)/13. \qquad \ldots\ldots(16)$$

If, therefore, $r>(12r_1+r_2)/13$ the deviation is positive, and if $r<(12r_1+r_2)/13$ the deviation is negative. The extent of the deviation is given by equation (14).

The systems under consideration may be discussed in turn:

(i) Silver in copper. The radius of the silver atom is 1·5933A., as compared with 1·4099A. for copper. The quantity $(12r_1+r_2)/13$ is thus 1·4240 A., which is less than the equilibrium value of the radius of atoms in the group of thirteen atoms. The deviation from Vegard's law is thus positive, as observed. The calculated value of $d\Delta\rho/dC$ is 0·021, as compared with the experimentally observed value of 0·020.

(ii) Copper in silver. In this case the quantity $(12r_1+r_2)/13$ is 1·5792 A. so that it is again less than the equilibrium value of the atomic radius within a group.

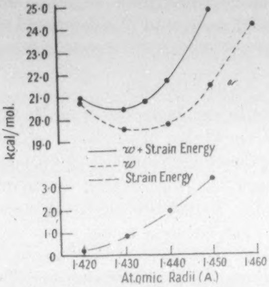


Figure 3. Copper–silver alloy system, copper-rich.

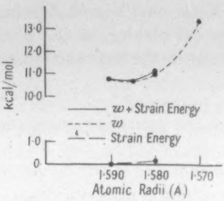


Figure 4. Copper–gold alloy system, gold-rich.

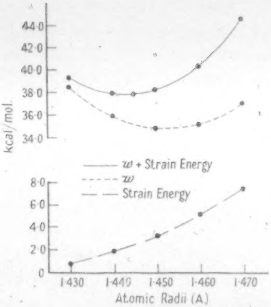


Figure 5. Copper–gold alloy system, copper-rich.

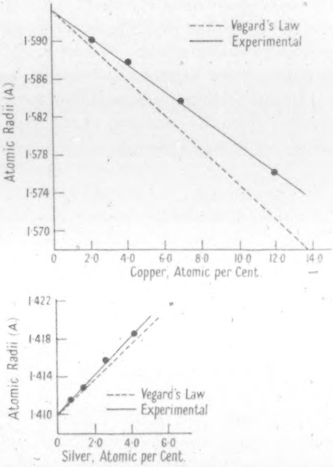


Figure 8. Copper–silver alloy system: deviations from Vegard's law (Owen and Rogers 1935).

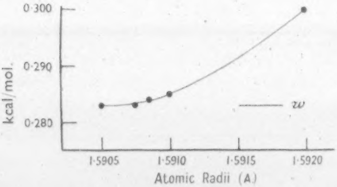


Figure 6. Gold–silver alloy system, gold-rich.

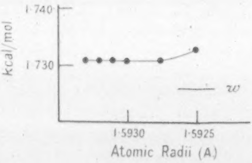


Figure 7. Gold–silver alloy system, silver-rich.

The deviation from Vegard's law is again positive, and the rate of change of the deviation with composition may be calculated as 0·076. This is not in such good agreement with the experimental quantity (0·042) as in the case of silver in copper, but is of the correct order of magnitude. It is of interest to note that the theory indicates correctly that the magnitude of the deviation from Vegard's law is greater for silver-rich alloys than for the copper-rich alloys.

Table 2

Alloy	r	$12\epsilon_2(r)+\epsilon_1(r)$	w	Strain Energy	r_{equilib}	Δ_{12}
Cu in Ag	1·590	−884·7	12·5	0·0		
	1·595	−885·2	12·0	0·04		
	1·580	−885·2	12·0	0·14	1·583	12·2
	1·570	−884·5	12·7	0·60		
	1·560	−882·4	14·8	1·30		
Ag in Cu	1·420	−1021·6	20·8	0·21		
	1·430	−1022·8	19·6	0·84		
	1·440	−1022·6	19·8	1·95	1·428	20·5
	1·450	−1020·9	21·5	3·35		
	1·460	−1018·2	24·2	—		
Cu in Au	1·5905	−1174·4	10·8	0·0		
	1·5850	−1174·5	10·7	0·03	1·587	10·7
	1·5800	−1174·2	11·0	0·14		
	1·5700	−1171·8	13·4	—		
Au in Cu	1·430	−1027·9	38·5	0·84		
	1·440	−1030·4	36·0	1·88		
	1·450	−1031·5	34·9	3·35	1·443	37·8
	1·460	−1031·2	35·2	5·22		
	1·470	−1029·3	37·1	7·51		
Ag in Au	1·59050	−1171·717	0·283	—		
	1·59075	−1171·717	0·283	—		
	1·59085	−1171·716	0·284	—	1·59065	0·3
	1·59100	−1171·715	0·285	0·0003		
	1·59200	−1171·700	0·300	—		
Au in Ag	1·59330	−906·269	1·731	—		
	1·59320	−906·269	1·731	—		
	1·59310	−906·269	1·731	—	1·59308	1·7
	1·59300	−906·269	1·731	—		
	1·59275	−906·269	1·731	0·0005		
	1·59250	−906·267	1·733	—		

Values of r are in A.; all other units are kcal/mol. $\Delta_{12}=w+$ strain energy at r_{equilib}.

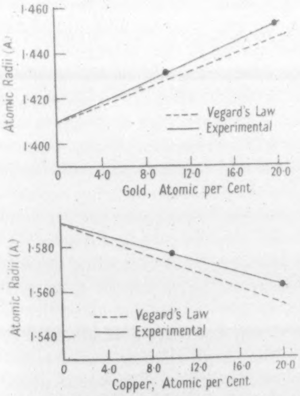


Figure 9. Copper–gold alloy system: deviations from Vegard's law (Van Arkel and Basart 1928).

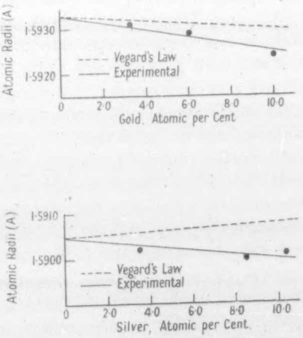


Figure 10. Gold–silver alloy system: deviations from Vegard's law (Sachs and Weerts 1930).

(iii) Gold in copper and copper in gold. It will be appreciated that from the curves of (w + strain energy) plotted against r it is difficult to locate accurately the exact position of the equilibrium values of r at the minima, owing to the shallow nature of the trough. For the copper-rich alloys the minimum occurs at $r = 1{\cdot}435$–$1{\cdot}446$ A., while the value for gold-rich alloys is $1{\cdot}585$–$1{\cdot}590$ A. The corresponding values of $(12r_1 + r_2)/13$ are $1{\cdot}4238$ A. and $1{\cdot}5766$ A., which are both less than the appropriate values of r. Deviations are again positive, in agreement with experiment. Quantitative agreement is not good, as the theory suggests a rate of change with composition roughly five times that observed. This is discussed below.

(iv) Gold in silver and silver in gold. In this case, it is even more difficult to locate the value of r at equilibrium for the atoms within a group, and quantitative comparisons are not justified. For the silver-rich alloys $(12r_1 + r_2)/13 = 1{\cdot}5931$ A., which is greater than the value of r_{equilib} in Table 2. Similarly, for the gold-rich alloys $(12r_1 + r_2)/13 = 1{\cdot}5907$ A. which is greater than r_{equilib}. In this case a negative deviation from Vegard's law is expected, and is observed, though the data, owing to the shallowness of the minima in Figures 6 and 7, are not critical.

The theory, therefore, gives the correct sign of the deviation from Vegard's law, and quantitative agreement is good for the copper–silver alloys.

It is of interest to consider qualitatively the implications of this treatment, remembering that the theory postulates that atoms comprising a group are equal in size, while any atom is more easily expanded than contracted, owing to the forms of the curves of ϵ against r. These curves show that copper and silver are more easily expanded and contracted than gold. Thus, when silver is added to copper, the silver atom resists contraction, and the equilibrium atomic radii of the polyhedra in a group are greater than $(12r_1 + r_2)/13$. When copper is added to silver, the solute atom expands easily almost to the size of the silver atom, and again a positive deviation is shown. Since an atom expands more easily than it will contract, the largest deviations from Vegard's law are to be expected on the side of the equilibrium diagram with the larger atom as solvent. This is experimentally true for the cases studied here. The most interesting comparison is between the copper–gold and silver–gold systems. Gold is both less easily expanded and contracted than either silver or copper. Lattice spacings of dilute solid solutions in gold will thus tend to approximate to that of pure gold. If a smaller atom (copper) is added as solute, the contraction will be less than if the additive rule were obeyed, and a positive deviation results. A larger atom, however, such as silver, will not increase the lattice spacings as much as if the additive rule were obeyed, and the deviation will be negative. Since expansion of the metal atom is easier than contraction, it would be expected that, with gold as solvent, the deviation would be greater when the gold atom is being contracted than when it is being expanded. Thus, the positive deviation for copper in gold should be greater than the negative deviation for silver in gold, as experimentally observed. When gold is the solute, we may again regard it as tending to remain constant in size, so that if it is added to silver, of larger lattice spacing, the contraction of the lattice spacing of silver is greater than if the additivity rule were obeyed, and a negative deviation results. If, however, gold is added to a solvent with a smaller atom, such as copper, the lattice spacing increases more rapidly than the linear rate, resulting in a positive deviation. The rate of positive deviation should again be greater than the negative deviation, as observed for gold in copper and gold in silver. It is also of interest that, as indicated by the theory, the smallest deviation in the silver–gold system occurs in the silver-rich alloys, where the influence of gold is at a minimum.

As stated above, the calculated positive deviations for the copper–gold system, though correct in sign, are too big by a factor of approximately five. It is possible that this discrepancy is due to the operation of the electrochemical factor, which may result in the atoms being drawn more closely together than the theory suggests, since gold is relatively electronegative compared with copper.

The above discussion indicates that Vegard's law is to be regarded as inapplicable to solid solutions, since the relative characteristics of the atoms have to be taken into account.

§5. APPLICATION TO ALLOY EQUILIBRIA

The free energy of a solid solution may be written

$$F_{ss_{12}}(C,T)=\epsilon_0+\int_0^T C_p\,dT+\int_0^T \frac{C_p}{T}\,dT+N\boldsymbol{k}T[C\ln C+(1-C)\ln(1-C)]. \quad .(17)$$

In considering differences in free energy between phases, the terms involving the specific heat may be neglected and (17) simplified to:

$$F_{ss_{12}}(C,T)=\epsilon_0+N\boldsymbol{k}T[C\ln C+(1-C)\ln(1-C)]. \quad \ldots\ldots(18)$$

Since the free energy of a mixture of the pure components is:

$$F_{pp}=(1-C)\epsilon_1+C\epsilon_2 \quad \ldots\ldots(19)$$

it is clear that the solid solution will only be stable relative to this mixture provided

$$C\Delta_{12}+N\boldsymbol{k}T\{C\ln C+(1-C)\ln(1-C)\}<0. \quad \ldots\ldots(20)$$

The point at which the free energy of the solid solution and the phase mixture are equal, i.e. the point at which the two curves cross, is given by

$$C_1\Delta_{12}+N\boldsymbol{k}T\{C_1\ln C_1+(1-C_1)\ln(1-C_1)\}=0 \quad \ldots\ldots(21)$$

and the corresponding composition is

$$C_1=\exp\{-(\Delta_{12}-\boldsymbol{R}T)/\boldsymbol{R}T\}. \quad \ldots\ldots(22)$$

It may be noted that the smaller is the value of Δ_{12}, the larger will be the value of C_1, at which the free energy curves for the solid solution and the phase mixture cross. In general, therefore, there will be a greater probability of the formation of extensive solid solutions the smaller are the values of Δ_{12} and of Δ_{21}. It is of interest to consider how a knowledge of the magnitude of the value of Δ, calculated as described above, may be used to indicate the probability of solid solution formation as opposed to the formation of a phase mixture. For the solid solutions of silver in gold, and of gold in silver, the values of Δ are small (Table 2). This, according to the ideas expressed above, indicates that wide solid solutions are probable at both ends of the binary diagram, and it is not surprising that complete solid solution formation is observed experimentally. It will be appreciated that if, for both solid solutions in a binary alloy, the critical composition C_1 is large, there is a distinct possibility that the two free energy curves for the solid solutions will join up smoothly as illustrated by the broken line in Figure 11 (*a*). In contrast, Δ_{12} is large for the solid solutions of copper in silver and silver in copper, indicating more restricted solid solution formation than in silver–gold alloys, as observed.

The case of the copper–gold alloys is of interest. In this case, Δ_{12} for copper in gold is less than for copper in silver, while Δ_{12} for gold in copper is greater than for silver in copper. Experimentally, it is found that complete solid solutions exist. This would suggest that if Δ_{12}, calculated in the units employed in this paper, is less than approximately eleven, there is a strong tendency to continuous solid solution formation. To illustrate this, Figure 11 (*b*) has been constructed to show how it might be possible for the continuous solid solution to be formed in such a case where one value of Δ is relatively small, and the other large. This possible criterion for complete solid solution formation will be more fully examined in a later contribution to this series, dealing with the equilibrium relationships in the alloys of the monovalent metals of Group IA of the Periodic Table.

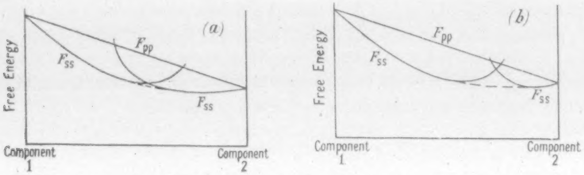


Figure 11. Diagrammatic representation of free energy curves: (*a*) Δ_{12} and Δ_{21} small; (*b*) Δ_{12} small, Δ_{21} relatively large.

The above treatment refers to dilute solutions, in which Δ_{12} may be assumed constant. In order to gain a fuller understanding of the solubility relationships it is desirable to extend the treatment to higher concentrations, when it is very probable that Δ_{12} will decrease with concentration of solute. The problem of the formation of phase boundaries will therefore be discussed in more detail in a subsequent publication.

§ 6. SUMMARY AND CONCLUSIONS

In order to extend the theory of the cohesion of metals to alloy structures, a new treatment is suggested, in which the excess energy of the solid solution over and above that of a mixture of the pure components is calculated by considering that the effects introduced by a solute atom cause a major disturbance extending only to its nearest neighbours and a less serious disturbance in the surrounding matrix. For a face-centred cubic solid solution, a group of thirteen atoms is thus formed by a solute atom and its twelve nearest neighbours. The size of the atomic polyhedra forming this group is different from that of the atomic polyhedra in the matrix, and the equilibrium size of the polyhedra within the group may be calculated from the known lattice energies of the components, the atomic radii of the components, and the strain energy introduced, which is a measure of the minor disturbance in the matrix, assuming that the variation with volume of the energies of the components in the alloy is the same as in the pure metals. From the size of the polyhedra, the deviations from Vegard's law to be expected in the solid solutions formed by copper, silver and gold, taken in pairs, may be evaluated, with qualitative agreement with experiment; quantitative agreement is obtained for the copper–silver alloys, which provide the most favourable case for treatment.

A quantity Δ is evaluated, which expresses the total energy possessed by the alloy over and above that of a mixture of the components. It is shown that this value is of considerable importance in the determination of whether or not phase boundaries occur in a given alloy system; although the theory is not yet complete in this connection a criterion is suggested. The treatment of the problem outlined in the paper may indicate the lines to be followed in a more formal treatment for a complete discussion of these systems.

ACKNOWLEDGMENTS

This work was carried out in the Metallurgy Department of the University of Birmingham, under the general direction of Professor D. Hanson, to whom the authors' thanks are due for his interest and support. Grateful acknowledgment is also made to the University in respect of a Fellowship held by one of us (M. A. J.) and to the University of Toronto in respect of the Raymond Priestley Fellowship held by a second (W. G. H.).

REFERENCES

ARAFA, M. K. I., 1949, *Proc. Phys. Soc.* B, **62,** 238.

FRENKEL, J., 1946, *Kinetic Theory of Liquids* (Oxford : University Press).

FUCHS, K., 1935, *Proc. Roy. Soc.* A, **151,** 585.

FÜRTH, R., 1944, *Proc. Roy. Soc.* A, **183,** 87.

JONES, H., 1937, *Proc. Phys. Soc.*, **49,** 243, 250.

MOTT, N. F., and JONES, H., 1936, *The Theory of the Properties of Metals and Alloys* (Oxford : University Press).

MOTT, N. F., and NABARRO, F. R. N., 1940, *Report on Internal Strains in Solids* (London : Physical Society), p. 86.

OWEN, E. A., and ROGERS, J., 1935, *J. Inst. Met.*, **42,** 257.

PINES, B. J., 1940, *J. Phys. U.S.S.R.*, **3,** 309.

SACHS, G., and WEERTS, J., 1930, *Z. Phys.*, **60,** 481.

SEITZ, F., 1935, *Phys. Rev.*, **47,** 400.

VAN ARKEL, A. E., and BASART, J., 1928, *Z. Kristallogr.*, **68,** 475.

WIGNER, E., 1934, *Phys. Rev.*, **46,** 1002.

WIGNER, E., and SEITZ, F., 1933, *Phys. Rev.*, **43,** 804; 1934, *Ibid.*, **46,** 509.

The Cohesion of Alloys: II
Some General Metallurgical Implications

By W. G. HENRY, M. A. JASWON* and G. V. RAYNOR
Metallurgy Department, University of Birmingham

Communicated by H. Jones; MS. received 18th May 1950, and in final form 11th October 1950

ABSTRACT. On the basis of the theory developed in the first paper of this series, certain phenomena of general theoretical interest may be interpreted. The theory indicates that, for cases of mutual solid solubility of the two components of a binary alloy, the wider solution should be formed where the smaller ion is dissolving in the solvent with the larger ion. This is in agreement with observation. The theory also interprets adequately the observation that, in general, the less compressible of a given pair of metals tends, at a given temperature, to dissolve more of the metal with the higher compressibility than the reverse.

The possible effect on solid solubility relationships of the mutual distortion of the electron clouds of the component ions in a binary alloy is also discussed from the point of view of the theory, and it is concluded that such distortion effects are insufficient to account for the difference between the equilibrium diagrams for the copper–gold and copper–silver alloys. It is suggested that mutual distortion of the ions is not an important effect for the alloys formed by metals in Group IB of the Periodic Table, taken in pairs. For alloys containing small ions bearing a high charge, however, this factor may still be of importance.

---

§ 1.

In Part I of this series (Jaswon, Henry and Raynor 1951), a theory has been outlined which enables the assessment of the cohesive energy of simple alloys of the metals of Group IB of the Periodic Table with each other, and has been applied in detail to the deviations in dilute solution from Vegard's law. In the present paper certain more general metallurgical implications of the theory are discussed.

In essence, the theoretical treatment is semi-empirical. The major disturbance caused by the introduction of a solute atom is considered to be limited to a region which includes only its nearest neighbours, so that attention is focused on groups of thirteen atoms consisting of the solute atom and its twelve nearest neighbours in the face-centred cubic crystal structure. A minor disturbance affecting the matrix outside this group is also considered. The equilibrium size of the polyhedra within a group will, in general, be different from that of the atom of the solvent metal, because of the disturbance of the regular periodic field of the latter. This equilibrium size may, however, be calculated on the assumption that the variation of binding energy with volume for each atom in a group remains the same as for the pure metal components, and that the minor disturbance introduced into the surrounding medium may be taken into account by considering the strain energy in the matrix around a group. It is thus possible to assess the total energy possessed by the alloy in excess of that which would be possessed by a phase mixture of the two components. This theory adequately interprets the signs of the observed deviations from Vegard's law in the copper–gold, copper–silver and silver–gold alloys, with quantitative agreement between calculated and experimental magnitudes for the particularly favourable case of the copper–silver system.

---

* Now at the Department of Mathematics, Imperial College, London.

§ 2.

For a composition C, the binding energy of a phase mixture of the pure components is clearly

$$\epsilon_p = (1-C)\epsilon_1 + C\epsilon_2, \qquad \ldots\ldots(1)$$

where ϵ_1 and ϵ_2 are the binding energies of the solvent and solute respectively; and the corresponding energy for a solid solution may be written

$$\epsilon_0 = (1-C)\epsilon_1 + C\epsilon_2 + C\Delta_{12}, \qquad \ldots\ldots(2)$$

so that $C\Delta_{12}$ represents the total energy of the solid solution over and above that possessed by a corresponding phase mixture.

The quantity Δ_{12} is made up of two separate parts which are, however, interdependent. It includes the strain energy considered to approximate to the long-range disturbances involved on alloying, and an energy value w which interprets the local disturbance of the periodic potential by the introduction of the solute and the consequent change in the potential energy of the electron. In practice, the quantity w is found to be the most important of these factors. Considering a dilute solution (C small), and referring to the group of thirteen atoms containing a solute atom at the centre,

$$w = \{12\epsilon_1(r) + \epsilon_2(r)\} - \{12\epsilon_1(r_1) + \epsilon_2(r_2)\}, \qquad \ldots\ldots(3)$$

where r_1 and r_2 are the atomic radii of the solvent and solute atoms and r is the atomic radius of atoms within the group. A method for evaluating the equilibrium value of r has been given in Part I; the corresponding value of w is therefore calculable.

As an example, the system copper–gold may be considered. It is found that for the gold-rich alloys the equilibrium value of r for the atoms within a group is 1·587 A., as compared with values of 1·5905 A. and 1·4099 A. for gold and copper respectively. For the copper-rich alloys the equilibrium value of r for atoms within a group is 1·443 A. According to this treatment, therefore, the solute atom is confined to a volume not greatly different from that associated with an undisturbed solvent ion. The volume of the solute ion is therefore altered to a greater extent than that of the solvent ion and the relative magnitude of w must be determined by the change in volume of the solute ion.

Experimental curves for the dependence of binding energy upon volume, which may be generally expressed as

$$\epsilon = -\frac{A}{r^m} + \frac{B}{r^n} \qquad \ldots\ldots(4)$$

where $n > m$ show that for a given amount of either expansion or contraction the change in energy is always the greater for contraction. Hence the addition of a given quantity of solute with a large ion to a solvent with a small ion causes the solid solution to become less stable (free energy increase) than the addition of the same quantity of the small ion to the large; in the first case, according to the present interpretation, the solute is contracted, while in the second case it is expanded. Thus, for the addition of silver to copper the value of w is of the order of 20 kcal/mol, whereas for the addition of copper to silver the value of w is only approximately 12 kcal/mol. For the same increase in free energy, therefore, more copper may be taken into solution than silver. The present theoretical treatment would thus indicate that, for cases of mutual solubility of the two components of a binary alloy, the wider solid solution should be formed in the case where the smaller ion is dissolving in the solvent with the larger ion. This is entirely in accord with the experimental results summarized by Hume-Rothery and Raynor (1938), who came to the conclusion that, in general, for a given pair of metals of the 'full' type, the solid solution in the metal with the larger ionic diameter was the greater at any given temperature. The present theory provides a quantitative interpretation of this generalization.

It should be pointed out that the above considerations imply that the coefficients m and n which determine the curvature of the semi-empirical energy curves for the component metals are not greatly different. If, for instance, the curvature of the appropriate curve for a small ion was such that the expansion of the small ion was more difficult than the contraction of a large ion, the relative size effect would not operate. The calculations of Fürth (1944), however, indicate that the values of m and n for various metals do not differ sufficiently widely for such a case to be probable.

The coefficients m and n are directly related to the compressibility of the metal by the following expression, which is due to Fürth:

$$mn \propto qV/K\Lambda, \qquad \ldots\ldots(5)$$

where V is the atomic volume, K the compressibility, and Λ the sublimation energy determined at room temperature. In general, therefore, the smaller the compressibility, the greater is the product of m and n and the more sharply defined is the minimum on the curve of energy against volume for a given metal. The case of the mutual solubility of two metals A and B of differing compressibilities may now be considered; A is assumed to have the higher compressibility. The curve of binding energy against the atomic radius has therefore a more shallow minimum than would be the case for metal B, and a given degree of either expansion or contraction will cause less of an increase in energy. It would be expected, therefore, that on dissolving a metal with a relatively high compressibility in one with a lower compressibility, the contribution to w will be less than in the reverse case. A greater amount of A would thus be required in B than of B in A to raise the free energy of the solid solution by a given amount, and the widest solid solution would be expected to exist in the metal with the lower compressibility. This again is in agreement with experiment. It has already been observed (Raynor 1938) that, in general, the less compressible of a given pair of metals tends, at a given temperature, to dissolve more of the metal with the higher compressibility than the reverse. This is especially true for metals of the full type. The application of the ideas of the present theory provides an adequate interpretation of the mutual solid solubilities of pairs of metals both in terms of the relative atomic sizes and the relative compressibilities.

§ 3.

Jaswon, Henry and Raynor (1951) have shown that the application of the present theory to the alloys of copper, silver and gold gave an adequate interpretation of the existence of a miscibility gap in the silver–copper alloys and of the greater solubility of gold in copper than of silver in copper. It was, however, not possible to interpret the complete solubility which exists between copper and gold, because in this case the quantity Δ_{12} is relatively smaller, and the concentrations involved are relatively larger, so that the approximation of dilute solution is no longer valid.

Several attempts have been made to account for the puzzling difference between the behaviour of these two alloy systems, in which such metallurgical factors as the atomic size factor, electronegative valency effect, and crystal structure are almost identical. Hume-Rothery and Raynor (1938) examined the anomaly in terms of the relative polarizing powers and polarizabilities of the copper, silver and gold ions, and suggested that, when metals are compared on solvents, other things being equal, those metals whose ions are the strongest polarizers would take up other metals more easily into solution, since the strong polarizing power of the solvent might be able to deform the charge distribution of the solute ion and permit it to fit more readily into the solvent lattice. The polarizing powers of the ions concerned were assessed, from the nature of the bond type in their halide salts, in the order $Cu^+ > Au^+ > Ag^+$, while the polarizabilities are given by Van Vleck (1932) in the order $Au^+ > Ag^+ > Cu^+$. With copper as solvent, gold is more easily deformed than silver, and wider solid solution is expected. With copper as solute, gold is a better polarizer than silver, so that the gold-rich solution might be expected to be the wider. Considerations involved in the development of the present theory, however, make it doubtful whether this is the real explanation.

In the first place, if polarization in the chemical sense occurs, some of the electron cloud of the solute atom may be considered to come partially under the control of the surrounding solvent ions. This leads to a tendency towards the establishment of electron-pair bonds, and hence a tendency towards a diminution of the internuclear distances as a result of the deformation process might be expected. In the copper–gold system, however, which is a combination of the strongest polarizer and the most easily polarized ion among the Group IB metals, and might thus be expected to show a negative deviation from Vegard's law, the observed deviations are positive. If the polarization factor is present in these alloys, therefore, its effect appears to be outweighed by the factors discussed in the development of the present theory, which successfully interprets the observed deviations. Similarly, according to the polarization hypothesis, negative deviations would be expected in the copper–silver system, instead of the observed positive deviations. In the silver–gold system, where the polarization effect would be expected to be least effective, and where, consequently, a marked negative deviation from Vegard's law would not be expected, experiment shows that a negative deviation is observed. The observed deviations from linearity in the lattice-spacing/composition relationships for the binary alloys of the Group IB metals thus do not support the polarization hypothesis as a major influence in alloy formation.

The tendency for the less compressible of a pair of metals to dissolve more of the metal of higher compressibility than the reverse has already been referred to. Since the polarization of an ion is a measure of the capacity of an ion to deform under the influence of electrical forces, and since the compressibility of a substance is a measure of its capacity to deform under external pressure, some connection between these two metallic characteristics might be expected. According to the published tables of Kaye and Laby (1948), the compressibilities of copper, silver and gold are: gold $0{\cdot}60 \times 10^{-6}$, copper $0{\cdot}74 \times 10^{-6}$, silver $0{\cdot}92 \times 10^{-6}$. Gold is thus the least compressible of the Group IB metals, and in the metallic state would appear to be the least easily polarized, contrary to the observations on ionic crystals.

In discussing the alloys of copper, silver and gold with each other, therefore, the conception of polarization in the chemical sense should be discarded as a major influence in their formation. Even in ionic crystals, the cation is considerably less deformed than the anion, since the electron cloud is held more rigidly by the excess positive charge than the electron cloud about the anion. In the case of the monovalent metals under discussion, which consist only of cations, and in which the ionic and atomic radii are comparable, the polarizing forces must be small, partly because of the single charge on the ions, and partly because the polarizing forces fall off as r^{-2}, where r is the ionic radius, which is relatively greater in metals than in ionic salts.

On the other hand, for alloys of metals which have small ions bearing a high charge, such as aluminium (Al^{+++}, $r = 0{\cdot}72$ A.), the concept of polarization is more legitimate and should still be considered as one of the metallurgical factors of importance in governing the details of the process of alloy formation (Raynor 1938).

§ 4.

In the present paper it has been shown that the conceptions of the theory of cohesion in the alloys of Group IB metals previously presented lead to generalizations which are in accord with experiment, particularly in connection with the influence upon the mutual solid solubilities of two metals of their relative atomic sizes and compressibilities. The tendency for the solid solution of the metal with the smaller ion in the metal with the larger ion to be more extensive than the reverse, and the tendency for a metal of low compressibility to dissolve more of the metal of higher compressibility than the reverse, are both quite general. It appears probable therefore that the theoretical conceptions developed are also capable of wider application than to the alloys of Group IB only.

In the development of the theory, no account is taken of any possibility that the free energy of the solid solution be lowered except by the entropy of mixing term in expression for the free energy. Δ_{12} (equation (2)) is always positive, so that it must be a condition of the treatment that the components of the alloy should be relatively inert towards each other. No account is taken, therefore, of possible strong electrochemical interaction between components, so that the presence of such a factor would limit the applicability of the theory. It would be of interest, however, to examine the application of the theory to the alloys with each other of the metals of Group IA of the Periodic Table, since these alloys form a relatively simple class in which disturbing factors are at a minimum.

The further discussion of the alloys of copper, silver and gold in terms of the ideas of the present theory make it appear probable that, in these alloys, polarization phenomena are not of major importance in determining the alloying characteristics.

ACKNOWLEDGMENTS

This reaserch was carried out in the Metallurgy Department of the University of Birmingham, under the general direction of Professor D. Hanson, to whom the authors' thanks are due for his interest and support. One of us (M. A. J.) expresses his gratitude to the University for a Research Fellowship, and another (W. G. H.) thanks the University for election to the Raymond Priestley Fellowship.

REFERENCES

Fürth, R., 1944, *Proc. Roy. Soc.* A, **183,** 87.

Hume-Rothery, W., and Raynor, G. V., 1938, *Phil. Mag.*, **26,** 143.

Jaswon, M. A., Henry, W. G., and Raynor, G. V., 1951, *Proc. Phys. Soc.* B, **64,** 177.

Kaye, G. W. C., and Laby, T. H., 1948, *Physical and Chemical Constants* (London: Longmans Green).

Raynor, G. V., 1938, *Phil. Mag.*, **26,** 152.

Van Vleck, J. H., 1932, *The Theory of Electric and Magnetic Susceptibilities* (Oxford: University Press), p. 225.

The Cohesion of Alloys: III
Extension to High Solute Concentrations, and Application to the Alloys of Group IA Metals

By W. G. HENRY and G. V. RAYNOR

Communicated by H. Jones; MS. received 18th May 1950, and in final form 11th October 1950

ABSTRACT. The theory developed by Jaswon, Henry and Raynor (1951) in the first paper of this series permits an estimate to be made, in dilute solution, of the extra energy Δ_{12} associated with the solution of metal 2 in metal 1 to form an alloy. In order to apply this theory to problems of solid solution formation, it is necessary to extend it to cover cases of relatively concentrated solution, when the variation of Δ_{12} with composition must be considered. A reasonable extension is discussed in the present paper, making the assumption that Δ_{12} varies linearly with composition. It is shown that if Δ_{12} is large, limited solid solution formation is to be expected, while if Δ_{12} is small, extensive or complete solid solution formation is possible. Calculations made for the alloys of the alkali metals, taken in pairs, give results in good agreement with experiment, and interpret the fact that solid solution formation is restricted in all the possible systems except K–Rb, K–Cs, and Rb–Cs, for which extensive solid solutions exist.

Further extension of the theory to include the liquid state enables the relative magnitudes of the liquidus depressions in the systems K–Rb, K–Cs, Rb–Cs, Na–K, Na–Rb and Na–Cs to be accounted for, and also interprets the compositions of the eutectic points in the systems Na–K, Na–Rb and Na–Cs. The formation of immiscible liquid systems is also discussed.

§1. INTRODUCTION

In the previous papers of this series (Jaswon, Henry and Raynor 1951, Henry, Jaswon and Raynor, 1951) to be referred to as I and II, consideration has been given to a method of assessing the total energy possessed by a binary alloy over and above that characteristic of a mixture of the two components. The theory has been applied to the deviations from Vegard's law shown in the systems copper–gold, silver–gold, and copper–silver, and some general metallurgical implications of the method have been discussed. In its simple form, however, the theory can be applied only to dilute solid solutions, and, for the purpose of using it to interpret solid solubility formation and the forms of metallurgical equilibrium diagrams, it is necessary to extend the treatment to cover more concentrated solid solutions. It is the purpose of the present paper to examine this problem.

The energy of a mixture of components 1 and 2, in which the concentration of metal 2 is C, is:

$$\epsilon_p = (1-C)\epsilon_1 + C\epsilon_2, \qquad \ldots\ldots(1)$$

where ϵ_1 and ϵ_2 are the cohesive energies of the pure components. For an alloy of the same composition, the cohesive energy

$$\epsilon_0 = (1-C)\epsilon_1 + C\epsilon_2 + C\Delta_{12} \qquad \ldots\ldots(2)$$

where Δ_{12} represents the increase in energy associated with the replacement of a solvent atom by a solute atom in forming a solid solution. In I a method for evaluating Δ_{12} was discussed, and it was shown that this quantity may be regarded as made up from the change in the potential energy of the electron, and a change in energy due to a longer range disturbance in the matrix round a solute atom. Of these, the change in the potential energy of the electron, w, is the most important.

The method of evaluating w involves the consideration of a solute atom surrounded, in the face-centred cubic structure, by a group of twelve solvent atoms. The primary disturbance is considered to be confined locally to this group, although a longer range disturbance is also present, and may be approximated by assessing the strain energy introduced into the matrix by the variation in volume of the twelve solvent atoms and one solute atom from that of thirteen solvent atoms. The change in potential energy may thus be written:

$$w=\{12\epsilon_1(r)+\epsilon_2(r)\}-\{12\epsilon_1(r_1)+\epsilon_2(r_2)\}, \qquad \ldots\ldots(3)$$

where ϵ_1 and ϵ_2 are respectively the cohesive energies of the components, r_1 and r_2 the atomic radii of the components, and r the radius of atoms within the group of thirteen which contains one solute atom. This treatment assumes that the variation of ϵ_1 and ϵ_2 with atomic radius is the same in the alloy as it is in the pure metals. Further, equation (3) can only be applied in cases where the concentration is smaller than a certain limit, below which the probability of finding two solute atoms within the group of thirteen atoms considered is negligible.

The quantity Δ_{12} may be regarded as:

$$\Delta_{12}=w+S, \qquad \ldots\ldots(4)$$

where S represents the strain energy introduced. Δ_{12} will only be independent of composition in very dilute solution.

§ 2. THE CONDITIONS FOR THE FORMATION OF PHASE BOUNDARIES

The free energy equations for N atoms of a phase mixture and a solid solution, both of composition C with respect to the solute 2, may be written as in equations (5) and (6). In these equations, the terms involving specific heat have been omitted, since we are interested only in a comparison of the two.

$$F_{\mathrm{pp}}=(1-C)\epsilon_1+C\epsilon_2 \qquad \ldots\ldots(5)$$

$$F_{\mathrm{ss}_{12}}(C,T)=(\epsilon_0+N\boldsymbol{k}T[C\ln C+(1-C)\ln(1-C)]. \qquad \ldots\ldots(6)$$

For dilute solution, equation (6) may be rewritten:

$$F_{\mathrm{ss}_{12}}(C,T)=(1-C)\epsilon_1+C\epsilon_2+C\Delta_{12}+N\boldsymbol{k}T[C\ln C+(1-C)\ln(1-C)]. \qquad \ldots\ldots(7)$$

The solid solution is stable relative to the phase mixture only if

$$C\Delta_{12}+N\boldsymbol{k}T[C\ln C+(1-C)\ln(1-C)]<0. \qquad \ldots\ldots(8)$$

Initially, at low solute concentrations, the curve of $F_{\mathrm{ss}_{12}}$ against composition lies below the corresponding curve of F_{pp} against composition, because the rate of change of $[C\ln C+(1-C)\ln(1-C)]$ at small values of C tends to $-\infty$. At some composition C_1, however, these curves may cross, and C_1 is given by the equation:

$$C_1\Delta_{12}+N\boldsymbol{k}T[C_1\ln C_1+(1-C_1)\ln(1-C_1)]=0. \qquad \ldots\ldots(9)$$

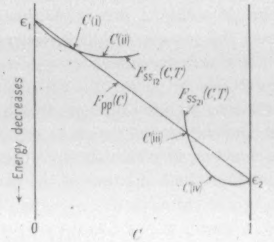
It is of interest to note that, assuming C to be small, the point at which the two curves cross is:

$$C_1=\exp\{-(\Delta_{12}-\boldsymbol{R}T)/\boldsymbol{R}T\}, \qquad \ldots\ldots(10)$$

whereas the minimum value of $F_{ss_{12}}(C, T)$, again assuming C to be small, lies at:

$$C = \exp\{-(\epsilon_2 - \epsilon_1 + \Delta_{12})/\boldsymbol{R}T\}. \qquad \ldots\ldots(11)$$

The minimum is only acceptable (i.e. at a concentration between $C=0$ and $C=1$) provided $(\epsilon_2-\epsilon_1+\Delta_{12})/\boldsymbol{R}T>0$. The values of both C_1 and C in the above equations depend upon the magnitude of Δ_{12}, and it is clear that the minimum value of $F_{ss_{12}}(C, T)$ may occur at lower or at higher concentrations than C_1 according to conditions. If $\epsilon_2<\epsilon_1$, then $(\epsilon_2-\epsilon_1+\Delta_{12})<\Delta_{12}$, neglecting $\boldsymbol{R}T$ since, for C to be small, the exponents in equations (10) and (11) must be large in comparison with $\boldsymbol{R}T$. Thus $C>C_1$ and the minimum occurs after the curves have crossed. If, on the other hand, $\epsilon_2>\epsilon_1$, the minimum occurs before the curves have crossed. This is illustrated in Figure 1.



$$C(\mathrm{i}) \quad C=\exp\left(\frac{\Delta_{12}-\boldsymbol{R}T}{\boldsymbol{R}T}\right); \qquad C(\mathrm{ii}) \quad C=\exp\left(\frac{\epsilon_2-\epsilon_1+\Delta_{12}}{\boldsymbol{R}T}\right);$$

$$C(\mathrm{iii}) \quad C=\exp\left(\frac{-(\Delta_{21}-\boldsymbol{R}T)}{\boldsymbol{R}T}\right); \qquad C(\mathrm{iv}) \quad C=\exp\left(\frac{-(\epsilon_1-\epsilon_2+\Delta_{21})}{\boldsymbol{R}T}\right).$$

Figure 1. Diagrammatic free energy curves for solid solutions and a mixture of the pure components.

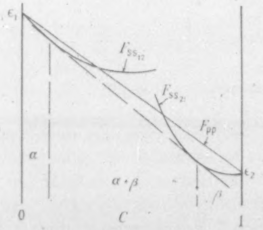


Figure 2. Diagrammatic free energy curves; Δ small, C_1 large.

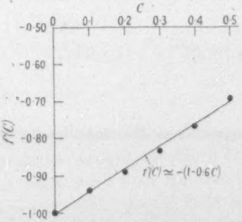


Figure 3. Approximation to $f(C)=\{(1-C)/C\}\ln(1-C)$. Circles denote values of this expression; the straight line is $f(C)=-(1-0{\cdot}6C)$.

The limits of the solid solubilities of the components in each other are defined by the points of contact of the common tangent to the curves of F_{ss} shown in Figure 1. It is clear that, the higher the value of C_1 at which the curves cross, the higher will tend to be the solute concentration at the point of contact of the common tangent, as indicated in Figure 2. Thus (equation (10)) the solid solubility must tend to rise as Δ_{12} decreases, so that restricted solid solutions will correspond to a higher value of Δ_{12} and more extensive solid solutions to a low value of Δ_{12}.

In evaluating Δ_{12}, however, and in deriving the solutions (10) and (11) it has been assumed that C is small, and that Δ_{12} is independent of composition (I). If Δ_{12} becomes itself small, then equations (10) and (11) are no longer accurate, and at the solubility limit it can no longer be assumed that Δ_{12} has the same value as in dilute solution.

It is necessary, therefore, to attempt to find a means of assessing the value of the concentration at which the curves against composition of F_{pp} and F_{ss} cross, without necessarily assuming C small.

§ 3. EXTENSION TO HIGHER CONCENTRATIONS

The quantity Δ_{12} is a measure of the positive energy associated with the introduction of one atom of solute 2 into a volume of solvent 1 which is sufficiently large to prevent the groups of thirteen atoms containing each solute atom from overlapping. It is therefore constant at low concentrations of solute only. At higher concentrations of solute Δ_{12} will decrease, since the foreign atom is being introduced into an environment which becomes progressively more like that in the solute metal itself. Δ_{12} must be recognized as a function of C, so that the limiting concentration at which the free energy curves for the solid solution and the corresponding phase mixture of the same composition cross is given by the condition:

$$C\Delta_{12}(C)+K[C\ln C+(1-C)\ln(1-C)]=0. \qquad \ldots\ldots(12)$$

In order to solve this equation, assumptions are necessary with regard to the variation of Δ_{12} with C, and the expression $(1-C)\ln(1-C)$ must be approximated.

Rewriting (12),
$$K\left[\ln C+\frac{1-C}{C}\ln(1-C)\right]=-\Delta_{12}(C). \qquad \ldots\ldots(13)$$

Table 1 and Figure 3 show that a reasonable approximation for

$$f(C)=\{(1-C)/C\}\ln(1-C)$$

is given by $f(C)=-(1-0{\cdot}6C)$.

Table 1

C	$\frac{1-C}{C}$	$\ln(1-C)$	$\frac{1-C}{C}\ln(1-C)$	$-(1-0{\cdot}6C)$
0·1	9·0	−0·104	−0·94	−0·94
0·2	4·0	−0·223	−0·89	−0·88
0·3	2·33	−0·358	−0·84	−0·82
0·4	1·50	−9·505	−0·76	−0·76
0·5	1·00	−0·694	−0·69	−0·70

For the variation of Δ_{12} with C it is assumed that:

$$\Delta_{12}(C)=a-bC, \qquad \ldots\ldots(14)$$

where a is the value of $\Delta_{12}(C)$ as C tends to zero and b is a constant. The equation (13) then simplifies to:

$$K[\ln C-(1-0{\cdot}6C)]=-(a-bC) \qquad \ldots\ldots(15)$$

or

$$C=\exp\{-(a-bC-K+0{\cdot}6KC)/K\}, \qquad \ldots\ldots(16)$$

where $K=N\boldsymbol{k}T$. This value of C gives the point of intersection of the two free energy curves.

The appropriate value of the constant b may be assessed, for the purposes of calculation, by finding the value of $\Delta_{0\cdot5}$ where the alloy consists of equal numbers of atoms of the two components. It has already been stated that the quantity w is the most important in deciding the magnitude of Δ_{12}, the strain energy factor being a second order quantity. We may thus write:

$$\Delta_{0\cdot5} \simeq w_{0\cdot5} \qquad \ldots\ldots(17)$$

and

$$w_{0\cdot5} = \{\epsilon_1(r) + \epsilon_2(r)\} - \{\epsilon_1(r_1) + \epsilon_2(r_2)\}, \qquad \ldots\ldots(18)$$

where $r = \frac{1}{2}(r_1 + r_2)$. b is therefore calculable to a first approximation. It is convenient to write

$$w_{0\cdot5} = \{E_1(r) + E_2(r)\} - \{\Lambda_1 + \Lambda_2\} \qquad \ldots\ldots(19)$$

where E_1 and E_2 are the binding energies for the pure components, expressed as a function of r, and Λ_1 and Λ_2 are the respective heats of sublimation $(\epsilon_1(r) = E_1(r) + E_{I_1};\ \epsilon_1(r_1) = \Lambda_1 + E_{I_1}$.

By the application of these methods, it is possible to estimate the composition at which the free energy curves for the solid solution and the phase mixture cross. If this occurs at a small value of C, conditions may be represented as in Figure 1, and the common tangent to the two curves indicates limited solid solubility. If, however, the compositions of the intersections, when evaluated from both ends of a binary system, are greater than 0·5, then extensive solid solutions are possible, as in Figure 4. In this case the free energy curves $F_{ss_{12}}(C, T)$ and $F_{ss_{21}}(C, T)$ may join up smoothly to give continuous solid solution formation (as indicated by the broken line in Figure 4), although it is difficult to predict this in any given case from the present treatment. It is, however, clear that the greater the value of C at which intersection occurs, the more extensive is the solid solution formation to be expected. If $C > 0\cdot5$ then continuous or at least very extensive solid solution formation is expected. In the next sections, the treatment outlined above is applied to the alloys of the metals in Group IA of the Periodic Table, and the results are compared with experiment.

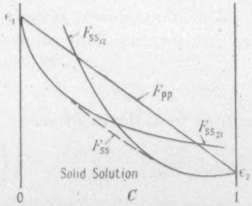


Figure 4. Diagrammatic free energy curves; $C_1 > 0\cdot5$.

§ 4. APPLICATION TO GROUP IA METALS

For the application of the theory outlined above to the alloys of the metals of Group IA of the Periodic Table, it is necessary to know the cohesive energies, the heats of sublimation, the atomic radii, and the variation of E_1 and E_2 with atomic radius. The variation of the binding energies E_1 and E_2 with r is considered to be given by the Fürth expression:

$$E_1(r) = \Lambda_1(r_1)\frac{nm}{n-m}\left\{-\frac{1}{m}\left(\frac{r_1}{r}\right)^m + \frac{1}{n}\left(\frac{r_1}{r}\right)^n\right\} \text{ per gm. atom.} \qquad \ldots\ldots(20)$$

Values of the atomic constants, and of m and n according to Fürth, are given in Table 2.

Table 2

Metal	m	n	$\Lambda_{18^\circ\text{C.}}$ (kcal.)	E_I (kcal.)	$\epsilon=\Lambda+E_\text{I}$ (kcal.)	r(A.)
Li	1·5	6	39·0	123·8	162·8	1·72
Na	2	6	25·9	118·0	143·9	2·10
K	2	6	19·8	99·7	119·5	2·61
Rb	2	6	18·9	95·9	114·8	2·81
Cs	2	6	18·8	89·4	108·2	3·04

According to equation (16), the composition of the intersection of the curves of F_{ss} and F_{pp} against composition is given by

$$C=\exp\{-(a-bC-K+0{\cdot}6KC)/\boldsymbol{R}T\}, \qquad \ldots\ldots(16)$$

where

$$\Delta_{12}(C)=a-bC=\Delta^0{}_{12}-bC \qquad \ldots\ldots(14)$$

and $K=N\boldsymbol{k}T$.

(i) *The Calculation of* $\Delta^0{}_{12}$

As explained above, the important factor in determining the value of $\Delta^0{}_{12}$ is the quantity $w^0{}_{12}$, and in this section $\Delta^0{}_{12}$ is taken as approximately equal to $w^0{}_{12}$, which must be calculated for a very dilute solution. When a very small amount of solute 2 is dissolved in solvent 1, which has the body-centred cubic structure, the application of the theory outlined in the first paper of this series indicates that the change in energy $w^0{}_{12}$ is given by

$$w^0{}_{12}=\{8E_1(r)+E_2(r)\}-\{8\Lambda_1+\Lambda_2\}. \qquad \ldots\ldots(21)$$

As shown in I and II, the solute atom tends to occupy a volume very similar to that of the solvent atoms, so that, for very dilute solution, we may without much loss of accuracy write:

$$w^0{}_{12}=\{8E_1(r_1)+E_2(r_1)\}-\{8\Lambda_1+\Lambda_2\}. \qquad \ldots\ldots(22)$$

The value of $E_2(r_1)$ is calculated from the expression:

$$E_2(r_1)=\Lambda_2(r_2)\,\frac{nm}{n-m}\left\{-\frac{1}{m}\left(\frac{r_2}{r_1}\right)^m+\frac{1}{n}\left(\frac{r_2}{r_1}\right)^n\right\}.$$

The values of $w^0{}_{12}$ and $w^0{}_{21}$ for the ten binary systems formed by taking the alkali metals in pairs are tabulated in Table 3. Also tabulated are the corresponding values for each system of $w_{0\cdot5}$, derived from equation (19).

Table 3

System 1 2	$\epsilon_2-\epsilon_1$ (k cal.)	Size factor $\frac{r_2-r_1}{r_2}\times100$ (%)	$w^0{}_{12}=\{8E_1(r)+E_2(r)\}-\{8\Lambda_1+\Lambda_2\}$ (kcal.)	$w^0{}_{21}=\{8E_2(r)+E_1(r)\}-\{8\Lambda_2+\Lambda_1\}$ (kcal.)	$w_{0\cdot5}=\{E_1(r)+E_2(r)\}-\{\Lambda_1+\Lambda_2\}$ (kcal.)
Li–Na	18·9	18·1	11·0	4·3	3·5
Na–K	24·4	19·6	11·0	4·3	3·1
K–Rb	4·7	7·2	0·7	0·5	0·3
Rb–Cs	6·6	7·6	1·0	0·5	0·3
Li–K	43·3	34·0	73·3	11·9	12·5
Na–Rb	29·1	25·0	21·1	6·4	5·2
K–Cs	11·3	14·1	3·9	1·8	1·4
Li–Rb	48·0	39·0	123·2	14·9	14·3
Na–Cs	35·7	31·0	46·2	8·3	9·1
Li–Cs	54·6	43·0	218·8	18·0	22·8

(ii) *The Calculation of the Constant b*

As shown above,

$$\Delta_{12}(C) = \Delta^0_{12} - bC \qquad \ldots\ldots(14)$$

so that b is simply $2(w^0_{12} - w_{0\cdot5})$ for the given case of solute 2 dissolved in solvent 1.

(iii) *The Calculation of the Critical Composition C*

The critical composition at which the free energy curves for the solid solution and the phase mixture cross may conveniently be derived graphically. As examples, Figure 5 and 6 give the constructions for the systems K–Cs and K–Rb respectively. For the former system, Table 3 shows that $a = 3{\cdot}9$ (K-rich) and $1{\cdot}8$ (Cs-rich), while $w_{0\cdot5}$ is $1{\cdot}4$. The values of b are therefore $5{\cdot}0$ (K-rich) and $0{\cdot}8$ (Cs-rich), and, for a temperature of 300° K., the equations for C become $C = e^{-(5\cdot5-7\cdot7C)}$ (solution of Cs in K), and $C = e^{-(2\cdot0-0\cdot73C)}$ (solution of K in Cs). In Figure 5, graphs are shown of the functions

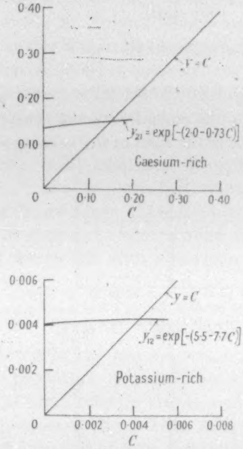


Figure 5. Construction for system K–Cs.

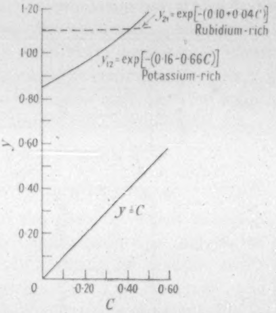


Figure 6. Construction for system K–Rb.

$$y = C \quad \text{and} \quad y = \exp[-(a - bC - K + 0{\cdot}6KC)/K];$$

the points of intersection give the critical values of C. For this system, it is seen that although at the K-rich end the critical value of C is small ($0{\cdot}0043$), that at the Cs-rich end is relatively large. Solid solution formation would thus be expected; in actual fact, the system contains uninterrupted solid solutions.

Figure 6 for the K–Rb system shows that the curves do not intersect at alloy concentrations below $0{\cdot}5$, so that in this case, uninterrupted solid solution formation would definitely be expected, as observed. In the case of the system Na–K, the intersections occur at very small concentrations. In this case no solubility would be expected, again as observed. Table 4 summarizes the critical concentration values for the ten systems under consideration.

Table 4

System 1 2	Critical Concentration (Soln. of 2 in 1)	Critical Concentration (Soln. of 1 in 2)	Type
Li–Na	small	small	III
Na–K	small	small	III
K–Rb	$>0{\cdot}5$	$>0{\cdot}5$	I
Rb–Cs	$>0{\cdot}5$	$>0{\cdot}5$	I
Li–K	very small	small	III
Na–Rb	very small	small	III
K–Cs	$4{\cdot}3\times10^{-3}$	$0{\cdot}15$	II
Li–Rb	very small	very small	III
Na–Cs	very small	very small	III
Li–Cs	very small	very small	III

small $= 2{\cdot}0\times10^{-3}$ to $1{\cdot}0\times10^{-9}$; very small $= <1{\cdot}0\times10^{-10}$.

§ 5. COMPARISON WITH EXPERIMENT

The equilibrium diagrams for the systems under consideration are summarized in Figure 7, and it is of interest to consider these in the light of the values contained in Table 4. It will be seen that the data in the Table tend to fall into three definite groups, as indicated in the Table. In group I (K–Rb and Rb–Cs) the critical concentrations, at which the free-energy/composition curves for the solid solutions and the mixture of pure components cross, exceed 0·5 for both ends of the binary system. Under these conditions, very wide solid solutions or more probably a continuous series of solid solutions would be expected, as discussed above. In agreement with this, potassium and rubidium are completely miscible in the solid state, and rubidium and caesium are similarly completely miscible. The third group, by contrast, contains systems in which the critical concentrations at which the relevant free energy curves cross are extremely small, and range from calculated values of $2{\cdot}0\times10^{-3}$ to less than $1{\cdot}0\times10^{-10}$. In these cases, no solid solution formation would be expected; the equilibrium diagrams are consistent with this explanation, and in no cases are extensive solid solutions found. In four of these cases (Li–Na, Li–K, Li–Rb and Li–Cs) immiscible liquids are also formed. Group II forms an intermediate group between the extremes already

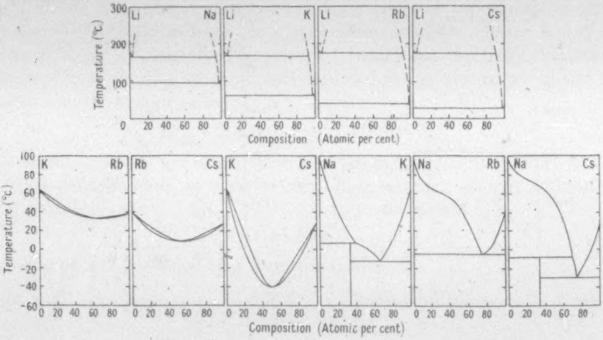


Figure 7. Equilibrium diagrams for the alloys formed by Group IA metals taken two together.

discussed. For the system K–Cs, it will be noted from Table 4 that the critical concentration for solution of potassium in caesium is appreciable (15 atomic per cent), so that wide solid solution formation would be expected. For the solution of caesium in potassium, however, the critical concentration is somewhat small, according to the theory developed in the previous sections of this paper. The equilibrium diagram of this system shows that a continuous series of solid solutions is formed, but it is of considerable interest to note that, whereas in the K–Rb and Rb–Cs systems, the liquidus and solidus curves coincide at a very shallow minimum, the corresponding minimum in the K–Cs system is pronounced. This may be interpreted as a tendency towards the splitting up of the solid solution into two phases. The system is therefore of a different type from the K–Rb and Rb–Cs systems, as would be expected from Table 4; the data in this Table, however, would not lead to the expectation of the complete miscibility observed in practice.

Comparison with experiment therefore confirms the general principle that in cases where the critical concentrations, at which the free-energy curves for the solid solutions and the phase mixture of pure components cross, are large, wide or continuous solid solution formation may be expected. For such cases $\Delta^0_{12}(\simeq w_{12}{}^0)$ and $\Delta^0_{21}(\simeq w^0_{21})$ are small, as previously discussed. Where the values of Δ are large, the corresponding critical concentrations are very small, leading to extremely limited solid solution formation such that the state of the system corresponds with a mixture of the almost pure components.

§6. FURTHER CONSIDERATIONS ON THE FORMS OF THE EQUILIBRIUM DIAGRAMS

As noted in the last section, certain of the alloys in which extensive solid solution formation is impossible show also immiscibility in the liquid state, while the system K–Cs shows a marked minimum in its liquidus and solidus curves in contrast to the systems K–Rb and Rb–Cs. This behaviour, and other features of the equilibrium diagrams, may be qualitatively understood in terms of the theory presented.

(i) *Cases where* $\Delta(C)$ *is small enough to allow complete solid solution formation.* We may consider free-energy composition curves $F_{ls}(C, T)$ for the liquid solution, and $F_{ss}(C, T)$ for the solid solution. As the temperature increases the free energy of both the liquid and the solid solution becomes lower, but owing to the relatively greater heat capacity of the liquid, the curves will approach one another, until at some temperature T',

$$F_{ls}(C', T') = F_{ss}(C', T') \qquad \ldots\ldots(23)$$

in which the values of the variables C' and T' correspond with the composition and temperature at which melting first occurs. At the given temperature T', this equality is only true for the composition C'.

As discussed above, the major factor in determining whether, in the solid state, an alloy system exists as a phase mixture or a solid solution is the value of $\Delta(C)$ $(\simeq w(C))$. It is therefore reasonable to suggest that this value will also determine the liquidus–solidus relationships. For the solid solution,

$$\begin{aligned} F_{ss_{12}}(C, T) = (1-C)\epsilon_1 + C\epsilon_2 + C\Delta_{12}(C) + K(C, T) \\ + N\boldsymbol{k}T[C\ln C + (1-C)\ln(1-C)], \qquad \ldots\ldots(24) \end{aligned}$$

while for the liquid mixture we may write

$$F_{\mathrm{ls}_{12}}(C,T)=(1-C)\epsilon_1+C\epsilon_2+C\theta_{12}(C)+K'(C,T)+L_{\mathrm{f}}(C)T/T_{\mathrm{m}}$$
$$+N\boldsymbol{k}T[C\ln C+(1-C)\ln(1-C)], \qquad \ldots\ldots(25)$$

where K and K' are functions of both temperature and composition. In equation (25), $\theta_{12}(C)$ is the positive energy associated with the solution of N atoms of metal 2 in metal 1 in the liquid state, and is a function of composition. L_{f} is the latent heat of fusion, T_{m} is the melting point, so that $L_{\mathrm{f}}/T_{\mathrm{m}}$ is the entropy of fusion, and is also a function of composition. The term $L_{\mathrm{f}}/T_{\mathrm{m}}$ is small, since for the metals considered here, L_{f} is of the order of only 500 cal/mol; to a first approximation it may be neglected. It is also assumed that the heat capacities of the liquid and solid solutions do not vary appreciably with concentration, and hence that their difference remains approximately constant.

In the solid state, the term $\Delta_{12}(C)$ is associated, as previously discussed, with the tendency of the solute atom to assume an effective size not far from that of the solvent atoms. In the liquid state the tendency to force a solute atom to conform to the size of its solvent atom neighbours will be much less than in a solid, since the atoms are much more loosely associated. It may be expected, therefore, that

$$\theta_{12}(C)<\Delta_{12}(C). \qquad \ldots\ldots(26)$$

For small values of $\Delta_{12}(C)$, we may take $\theta_{12}(C)$ as a second order quantity which may be neglected to a first approximation. If this is done, the term $C\Delta_{12}(C,T)$ in equation (24) determines the differences in curvature between $F_{\mathrm{ss}_{12}}(C,T)$ and $F_{\mathrm{ls}_{12}}(C,T)$, and hence the liquidus–solidus relationships.

When $\Delta_{12}(C)$ and $\Delta_{21}(C)$ are small, the form of $F_{\mathrm{ss}}(C,T)$ at a given temperature will be governed mainly by the entropy of mixing term, and the curve for the liquid will be similar (Figure 8, full lines). When the $\Delta(C)$ values are less small, the curve of $F_{\mathrm{ss}}(C,T)$ will be modified by a positive addition $C\Delta(C)$, with $\Delta(C)$ decreasing as C rises, and will be as shown by the broken line. As the temperature is increased the two curves approach, and touch at the temperature at which melting first occurs. Further increase of temperature causes the appearance of (solid + liquid) equilibria, represented by common tangents to the two curves. If the two curves are of similar curvature (full lines in Figure 8), then the transition from completely solid alloys to completely liquid alloys will occupy only a small temperature range, and the liquidus depression will be small. If, on the other hand, the curvatures differ appreciably (broken lines in Figure 8) then the temperature range occupied by the transition from the solid to the liquid

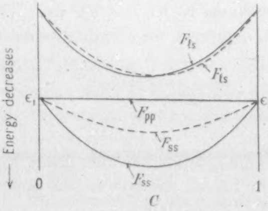


Figure 8. Diagrammatic free energy curves for solid and liquid solutions; full lines, Δ small; broken lines, Δ increasing.

state is increased; the liquidus depression is therefore greater than for a small value of $\Delta(C)$. As shown in Table 5 these considerations are in agreement with experiment. Where the $\Delta(C)$ values are small, the depression of the liquidus is relatively small, but for the system K–Cs, with larger $\Delta(C)$ values, the depression is twice as much as for the system Rb–Cs.

Table 5

System	1 2 K–Rb	1 2 Rb–Cs	1 2 K–Cs
w^0_{12}	0·7	1·0	3·9
w^0_{21}	0·5	0·5	1·8
$w_{0\cdot5}$	0·3	0·3	1·4
Maximum depression of liquidus [a] (°c.)	35	40	80
Position of minimum [b]	0·75	0·40	0·50
Liquidus–solidus separation	small	small	large

[a] Measured vertically from $(1-C)\,T_m(1)+CT_m(2)$ where C is the concentration of metal 2 at the minimum and $T_m(1)$ and $T_m(2)$ are the melting points of the pure components.

[b] Measured from the side of the diagram corresponding to metal 1.

At any given temperature in the range in which solid and liquid are in equilibrium, the limits of the two-phase regions are defined by the points of contact of the common tangents to the appropriate free-energy curves (Figure 9). It is easy to see that the more similar are the curvatures of these curves ($\Delta(C)$ small), the narrower will the two-phase regions be, whereas a more marked difference in curvature ($\Delta(C)$ less small) leads to a broadening of these regions. This is agreement with the equilibrium diagrams, in which the liquidus–solidus separation is greater for the system K–Cs than for the systems K–Rb and Rb–Cs.

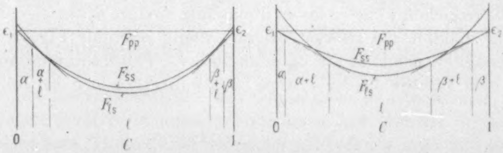


Figure 9. Dependence of width of (solid + liquid) field on relative curvature of free energy curves.

(ii) *Cases where $\Delta(C)$ is too large to allow solid solution formation.* Where $\Delta(C)$ is too large to allow solid solution formation, the liquid comes into equilibrium with a mixture of phases. For the Na–Rb alloys, the free-energy curves may be regarded as disposed as in Figure 10, and on approach, the curve $F_{ls}(C, T)$ will make contact with the straight line $F_{pp}(C, T)$. For the Na–K and Na–Cs systems, where intermetallic compounds Na_2K and Na_2Cs are formed, the corresponding diagram is as shown in Figure 11. Since these compounds

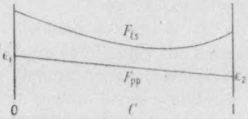


Figure 10. Diagrammatic free energy curves for solid and liquid solutions; Δ large.

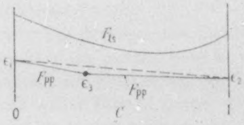


Figure 11. Diagrammatic free energy curves for solid and liquid solutions in the presence of a compound (e.g. Na_2K and Na_2Cs).

are 'size-factor' compounds of low heat of formation, it is likely that the point of intersection at the composition of the compound lies only slightly below the dotted line. For both types of diagram, alteration of temperature from that corresponding to contact of the curves leads to solid–liquid equilibria, and it is apparent that the temperature interval over which the transition from completely solid to completely liquid alloys occurs will be larger than that for the alloys considered above, which form continuous solid solutions. The very pronounced depression of freezing point in the systems Na–K, Na–Rb, and Na–Cs is thus accounted for. The position of the minimum on the liquidus (the eutectic point) is also accounted for by the following considerations. Where $\Delta(C)$ is relatively large, $\theta(C)$ must also be taken as of significance. The curve $F_{ls}(C, T)$ will be relatively lower on the side of the diagram corresponding with the smaller $\Delta(C)$ value, and hence the smaller $\theta(C)$ value. Contact will thus first occur on this side, determining the beginning of melting. The eutectic in the systems Na–K, Na–Rb and Na–Cs would thus be expected to lie on the side of the diagram remote from sodium (see Table 3). This is in agreement with experiment. It should be noted that, as discussed previously, the small values of $\Delta(C)$ (and hence of $\theta(C)$) occur when a smaller atom is dissolved in a solvent with a larger atom, because it is easier to expand an atom than to cause it to contract (II).

From Table 3, it may be noted that the systems Li–K, Li–Rb, and Li–Cs, which form immiscible liquids, have by far the highest values of $\Delta(C)(\sim w^0(C))$; the system Li–Na has values identical with those for the system Na–K. For the three first-mentioned systems, $\theta(C)$ must also be regarded as large. In these cases conditions in the liquid state are similar to those discussed in relation to the solid state in the earlier sections of this paper. Considering the relationship between the free-energy curves for an immiscible mixture of the pure liquids and for a liquid solution, the relevant diagram would resemble Figure 1, and the preferred condition would be that of immiscibility. This interpretation, however, does not hold for the Li–Na system, and must therefore be considered as incomplete.

ACKNOWLEDGMENTS

This research was carried out in the Metallurgy Department of the University of Birmingham, under the general direction of Professor D. Hanson, to whom the authors' thanks are due for his interest and support. One of us (W. G. H.) thanks the University for election to the Raymond Priestley Fellowship.

REFERENCES

Henry, W. G., Jaswon, M. A., and Raynor, G. V., 1951, *Proc. Phys. Soc.* B, **64**, 190.
Jaswon, M. A., Henry, W. G., and Raynor, G. V., 1951, *Proc. Phys. Soc.* B, **64**, 177.

The Thermal Conductivity of a Copper–Nickel Alloy at Low Temperatures

By J. K. HULM *

Royal Society Mond Laboratory, Cambridge

Communicated by D. Shoenberg; MS. received 16th October 1950

ABSTRACT. The thermal and electrical conductivities of an alloy of 80% Cu and 20% Ni with average crystal grain size 0·011 mm. have been measured at liquid hydrogen and liquid helium temperatures. A comparison of the results with Makinson's theory suggests that the thermal conductivity between 2° K. and 20° K. is the sum of a temperature proportional term due to electron transport with impurity scattering and a T^2 proportional term due to lattice wave transport with electron scattering, both terms being of the same order of magnitude.

§ 1. INTRODUCTION

THE experiments of Lees (1908) and Bremmer and de Haas (1932, 1936) were the first to reveal that when certain metallic alloys are cooled below room temperature, the thermal conductivity decreases slowly down to liquid nitrogen temperatures and then more rapidly as liquid hydrogen and liquid helium temperatures are approached. More recent measurements by Karweil and Schäfer (1939), Allen and Mendoza (1948) and Wilkinson and Wilks (1949) have indicated that a drop in thermal conductivity at low temperatures is a characteristic property of a fairly wide range of alloys. An explanation of this general effect is contained in Makinson's (1938) theoretical paper on heat transfer in metals and alloys, but the experimental data so far available are inadequate for a satisfactory test of the detailed predictions of the theory. In view of the need for more accurate, wide temperature range data on alloys of precisely known internal structure, and also in order to determine an important correction factor involved in some experiments on heat transfer in pure metals, described elsewhere (Hulm 1950) the author has recently measured the thermal and electrical conductivities of an alloy of 80% by weight copper and 20% nickel at liquid hydrogen and liquid helium temperatures. In the present paper, the results are compared with existing low temperature data for other similar alloys, and are discussed in the light of Makinson's theory and available knowledge on the internal structure of the alloy. Since alloys of this type are used extensively in the construction of cryogenic apparatus, a knowledge of factors influencing the thermal conductivity is worth while from the standpoint of experimental techniques as well as from the point of view of solid state theory.

§ 2. EXPERIMENTAL DETAILS

Since a detailed account of the experimental arrangement has been given elsewhere (Hulm 1950), only the main features of the method need be mentioned here. The determination of thermal conductivity was carried out by passing a known heat current along a uniform, thick-walled tube of copper–nickel alloy, and measuring the temperature difference between two points on the tube with a pair of helium gas thermometers. The gas thermometers were connected differentially, the pressure difference between them being measured on a butyl phthalate oil manometer read with a travelling microscope. With this arrangement temperature differences less than 0·02° K. could be measured with an accuracy within about 1% for specimen temperatures of a few degrees Kelvin, thus allowing a greater sensitivity and accuracy of measurement than in previous experiments on alloys.

---

* Now at the Institute for the Study of Metals, The University of Chicago.

The specimen was a tube of 5 mm. external and 3 mm. internal diameter selected at random from a long length of tubing supplied by Messrs. Johnson, Matthey Ltd., and fixed in position in the apparatus with lead solder. On completion of the conductivity measurements, a number of longitudinal and transverse cuts were made through the tube and the average crystal grain size was determined by electrolytically polishing, etching and examining under a microscope the surfaces thus exposed. A count of several hundred grains in both radial and axial directions gave practically the same average grain length of 0·011 mm. in each direction, so that in respect of grain size the specimen may be regarded as substantially isotropic.

§ 3. RESULTS

Experimental values of the thermal conductivity K of $Cu_{80}Ni_{20}$ at various temperatures in the liquid hydrogen and liquid helium ranges are set out in the Table and have been used to construct curve A of Figure 1 (where some of the liquid helium points are omitted to avoid crowding). A striking feature is the rapid drop in K with decreasing temperature, curve A having the same general shape as most of the remaining curves in Figure 1, which represent typical data obtained by previous workers for other similar alloys.

T (° K.)	K watt cm^{-1} deg^{-1} $\times 10^{-3}$	T (° K.)	K watt cm^{-1} deg^{-1} $\times 10^{-3}$	T (° K.)	K watt cm^{-1} deg^{-1} $\times 10^{-3}$
21·9	127	4·05	7·84	3·02	5·49
16·3	79·9	3·88	7·25	2·86	4·85
4·45	9·08	3·60	6·77	2·40	3·72
4·23	8·41	3·38	5·90	1·89	2·81

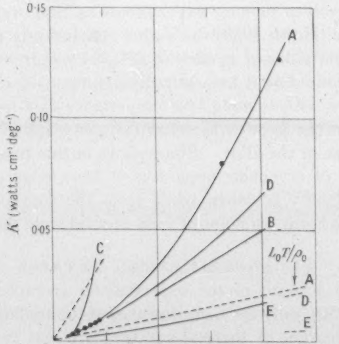


Figure 1. Thermal conductivity (full lines) and L_0T/ρ_0 (dotted lines) for various alloys at low temperatures.

A. $Cu_{80}Ni_{20}$, present work.
B. $Cu_{20}Ni_{30}$, Wilkinson and Wilks (1949).
C. $Cu_{46}Ni_{10}Zn_{42}Pb_2$, Allen and Mendoza (1948).
D. $Cu_{46}Ni_{13}Zn_{41}$, Karweil and Schäfer (1939).
E. $Ni_{60}Cr_{15}Fe_{16}Mo_7$, Karweil and Schäfer (1939).

The specific electrical resistance ρ of $Cu_{80}Ni_{20}$ changed by only a small amount from room temperature down to liquid hydrogen temperatures, as shown by the values: 289° K., $2{\cdot}59\times10^{-5}$ ohm cm.; 79° K., $2{\cdot}39\times10^{-5}$ ohm cm.; 20·4° K., $2{\cdot}33\times10^{-5}$ ohm cm.

It was therefore concluded that the residual resistance ρ_0 at the absolute zero lay within about 1% of $2{\cdot}33\times10^{-5}$ ohm cm.

§ 4. DISCUSSION

Although Koenigsberger (1907) pointed out as long ago as 1907 that the heat current in the interior of a metal subjected to a temperature gradient probably consists of the sum of separate currents carried by the conduction electrons and the lattice waves respectively, our knowledge of the relative importance of these two conduction processes as a function of such parameters as temperature, crystal grain size and impurity content or alloy composition remained for many years chiefly a matter of theoretical conjecture supported by comparatively little reliable experimental evidence, especially at low temperatures. Recent experiments (Hulm 1950) have indicated, however, that for metals with impurity contents less than about 0·1% the electronic thermal conductivity K_e at temperatures of a few degrees Kelvin is so large that the lattice thermal conductivity K_g may be neglected in comparison with it. On the other hand, an increase of impurity content above 0·1% reduces K_e to such an extent that the lattice waves contribute an appreciable part of the total heat current, and for specimens where the amount of impurity is comparable with that of the chief constituent, K_g may be expected to be of the same order of magnitude as or even greater than K_e. To meet this situation in the $Cu_{80}Ni_{20}$ specimen, it is convenient to use a method proposed by Gruneisen and Reddemann (1934) and Makinson (1938) to determine K_e from the electrical resistance of the specimen, thus allowing K_g to be studied separately.

At temperatures less than one-tenth of the Debye temperature Θ, Makinson's theoretical expression for the electronic thermal resistance of a metal may be written to a good approximation in the form:

$$1/K_e=\alpha T^2+\beta/T, \qquad \ldots\ldots(1)$$

where the two contributions arise, in order, from scattering of electrons by the vibrating ionic lattice and by impurity atoms. α is a constant for a given metal and β denotes ρ_0/L_0, where ρ_0 is the residual electrical resistance and L_0 is the Lorenz constant $\frac{1}{3}(\pi \boldsymbol{k}/\boldsymbol{e})^2$. Although α may be derived theoretically, it is perhaps more realistic to take the experimental value $2{\cdot}3\times10^{-5}$ watt^{-1} cm. deg^{-1} obtained by Gruneisen and Goens (1927) for copper. From the experimental value of ρ_0 for $Cu_{80}Ni_{20}$ we find that αT^2 is about 0·02% of β/T at 20° K., which indicates that the ionic lattice scattering term is negligible in the range of the present experiments. It is reasonable, therefore, to write the total thermal conductivity of $Cu_{80}Ni_{20}$ as:

$$K=K_g+L_0T/\rho_0, \qquad \ldots\ldots(2)$$

which may be used in conjunction with the data given in the Table to derive K_g at various temperatures. This procedure is only possible becaue L_0T/ρ_0 is always less than K for $Cu_{80}Ni_{20}$, a condition which according to Figure 1 is also satisfied for the specimens of Karweil and Schäfer, but not for the German silver specimen of Allen and Mendoza (curve C). Since for the latter specimen the dotted line L_0T/ρ_0 lies above the total conductivity curve over most of the measuring range, and in addition K seems to rise much more rapidly with increasing temperature than for any other alloy specimen, including Karweil and Schäfer's specimen of almost identical composition (curve D), some doubts are raised concerning the validity of curve C.

The values of K_g for $Cu_{80}Ni_{20}$ are plotted against temperature on logarithmic scales in Figure 2, where the fairly convincing straight line behaviour leads to the empirical law:

$$K_g = 2{\cdot}2 \times 10^{-4} T^{2{\cdot}0} \text{ watt cm}^{-1} \text{ deg}^{-1}. \qquad \ldots\ldots(3)$$

This result may be interpreted with the aid of Makinson's (1938) theory of lattice conduction, in which the following lattice wave scattering processes are considered: (i) scattering by electrons, (ii) scattering by crystal boundaries, (iii) scattering by impurity atoms and small-scale lattice defects, (iv) mutual scattering of the waves due to anharmonic terms in the equation of motion.

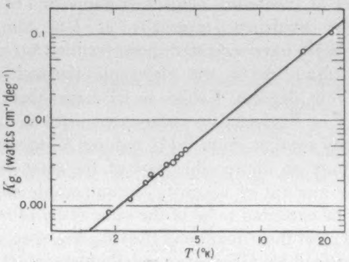


Figure 2. Temperature variation of the lattice conductivity K_g of $Cu_{80}Ni_{20}$ (log. scales).

In contributing to the total lattice conductivity these various processes tend to act as thermal resistances in series, so that the highest individual resistance or the lowest individual conductivity is predominant. Since theory indicates that the temperature coefficient of each individual conductivity is positive for the first two processes and negative for the last two processes, the total lattice conductivity–temperature curve passes through a maximum, below which electron and crystal boundary scattering are dominant and above which impurity scattering and mutual scattering are dominant. Equation (3) suggests that the maximum in K_g for $Cu_{80}Ni_{20}$ lies at a temperature above the range of the present measurements, a view which is further supported by theoretical estimates of the individual lattice conductivities for electron scattering and crystal boundary scattering in this specimen. Assuming one electron per atom, using the known crystal grain size and taking the Θ value 335° K. obtained by Keesom and Kurrelmeyer (1940) in low temperature specific heat measurements on a $Cu_{80}Ni_{20}$ specimen, we obtain from Makinson's equations (27) and (30):

$$\text{Scattering process (i), } K_g = 1{\cdot}1 \times 10^{-4} T^2 \text{ watt cm}^{-1} \text{ deg}^{-1}. \qquad \ldots\ldots(4)$$

$$\text{Scattering process (ii), } K_g = 3{\cdot}6 \times 10^{-4} T^3 \text{ watt cm}^{-1} \text{ deg}^{-1}. \qquad \ldots\ldots(5)$$

Since the coefficients but not the exponents of T are approximate in these formulae, equation (4) agrees quite well with our empirical equation (3), in support of the view that electron scattering is dominant in $Cu_{80}Ni_{20}$. The contributions of impurity scattering and mutual scattering appear to be negligible, while since K_g in equation (5) ranges from about 3 times to 300 times the observed lattice conductivity as T rises from 2° to 20° K., the effect of crystal boundary scattering is probably small.

It is worth mentioning in conclusion that equation (2) has been used to derive K_g for three of Karweil and Schäfer's specimens, the only other alloys beside $Cu_{80}Ni_{20}$ for which both K and ρ_0 are known and the condition $L_0T/\rho_0 \leqslant K$ is satisfied. The results are approximately represented by the formulae:

$Cu_{64}Ni_{16}Zn_{20}$, $K_g = 3{\cdot}8 \times 10^{-4} T^{2{\cdot}0}$ watt cm^{-1} deg^{-1}
$Cu_{46}Ni_{13}Zn_{41}$, $K_g = 4{\cdot}7 \times 10^{-4} T^{1{\cdot}5}$ watt cm^{-1} deg^{-1}
$Ni_{60}Cr_{15}Fe_{16}Mo_{7}$, $K_g = 2{\cdot}0 \times 10^{-4} T^{1{\cdot}4}$ watt cm^{-1} deg^{-1}.

The exponents of T suggest that electron scattering is dominant in the first specimen, but that a conductivity component proportional to T raised to a power less than 2 is important in addition to the electron scattering component in the last two specimens. This additional component is probably the conductivity due to impurity scattering (K_g proportional to T^{-1}) rather than that due to mutual scattering (K_g proportional to $\exp\{-\Theta/2T\}$), since the latter component is a too rapidly varying function of T to give a total lattice conductivity of the type observed for the last two specimens. It is interesting to note that the remarkably low thermal conductivity of $Ni_{60}Cr_{15}Fe_{16}Mo_{7}$, Contracid, arises from low values of K_e and K_g, both of these being closely related to the unusually high electrical resistivity for which nickel–chromium alloys are of technological importance at high temperatures. The production of alloys with even higher room temperature electrical resistivity than Contracid may provide a simple method of obtaining still better low temperature heat insulators with the desirable mechanical properties of metals.

ACKNOWLEDGMENTS

The author wishes to thank Mr. R. G. Chambers for measuring the electrical resistivity of the specimen and Mr. K. Ikeuye (Institute for the Study of Metals, University of Chicago) for assisting in the grain size determination. The main part of the work was carried out during tenure of a grant from the Department of Scientific and Industrial Research.

REFERENCES

Allen, J. F., and Mendoza, E., 1948, *Proc. Camb. Phil. Soc.*, **44**, 280.
Bremmer, H., and de Haas, W. J., 1936, *Physica*, **3**, 692.
Gruneisen, E., and Goens, E., 1927, *Z. Phys.*, **44**, 615.
Gruneisen, E., and Reddemann, H., 1934, *Ann. Phys., Lpz.*, **20**, 843.
de Haas, W. J., and Bremmer, H., 1932, *Commun. Phys. Lab. Univ. Leiden*, **220**, 12.
Hulm, J. K., 1950, *Proc. Roy. Soc.* A., **204**, 98.
Karweil, J., and Schäfer, K., 1939, *Ann. Phys., Lpz.*, **36**, 567.
Keesom, W. H., and Kurrelmeyer, B., 1940, *Physica*, **7**, 1003.
Koenigsberger, J., 1907, *Phys. Z.*, **8**, 237.
Lees, C. H., 1908, *Phil. Trans. Roy. Soc.* A, **208**, 381.
Makinson, R. E. B., 1938, *Proc. Camb. Phil. Soc.*, **34**, 474.
Wilkinson, K. R., and Wilks, J., 1949, *J. Sci. Instrum.*, **26**, 19.

Velocities of Steel Fragments after Perforation of Steel Plate

By K. E. SPELLS

Armament Research Establishment, Ministry of Supply*

MS. received 21*st August* 1950

ABSTRACT. Mild steel fragments (right square prisms, 1·593 cm. × 1·593 cm. × 0·711 cm.) were projected normally with velocity 1,469 m/sec. against mild steel plates of thickness between 0·21 cm. and 0·85 cm.; usually impact was with the full square fragment face. Perforation occurred by removal of a plug rather larger than the fragment face, but tending to diminish with plate thickness. Fragment and plug were shattered by the impact, the pieces emerging along somewhat divergent paths. Velocities were measured with a Cooper chronograph, and in general were approximately the same for the principal pieces of fragment and plug. The residual velocity of the principal pieces may be predicted with reasonable accuracy for plate thicknesses up to about 0·5 cm. by assuming momentum to be conserved between the attacking fragment and a plug of equal area. For thicker plates this leads to too high values of the residual velocity, presumably because of the neglect of the motion communicated to the rest of the plate.

Fragments of velocity 900 m/sec. suffered very little breakage against 0·21 cm. plates; residual velocities were not measured.

The mechanism of penetration is discussed.

§ 1. INTRODUCTION

During the recent war it was found necessary to measure the residual velocities of high speed mild steel fragments after passage through mild steel plates. As these experiments were only a subsidiary part of another investigation, their scope was limited to the projection of fragments of uniform size and shape (right square prism) with a velocity of 1,469 m/sec. against plates ranging in thickness from 0·21 cm. to 0·85 cm.

Few measurements of this type appear to have been published, however. From time to time, over a period of many years, measurements of the penetration of projectiles into different materials have been made, and some of the earlier work was summarized by Bashforth (1873), and by Resal (1895). Data are available for the velocities of bullets after passage through water (Gildmeister and Strehl 1905), and after passage through steel plates (Honda, Takermaé and Watanabe 1930). Measurements of the retardation of steel spheres of incident velocity between 150 and 650 m/sec. by a layer of water have also been made by Bauer (1926), but the perforation of steel plate by projectiles of simple geometrical shape does not seem to have been studied.

The determination of the residual fragment velocity was complicated by the break-up of the fragments which occurred during impact, and by the fact that the resulting pieces, together with the pieces of plate carried away, followed somewhat divergent paths. Only the velocity measurements were considered, and the investigation of the break-up was not pursued.

§ 2. CONDITIONS AND METHOD OF EXPERIMENT

The fragments used (1·593 cm. × 1·593 cm. × 0·711 cm.) were prepared by cutting slices perpendicular to the axis of a bright mild steel rod of 1·593 cm. square cross section and then grinding carefully to a thickness of 0·711 cm. They had a mass of 14·05 ± 0·07 gm., and were projected by the detonation of a small charge of high explosive, using a technique developed previously by members of the Research Department, Royal Arsenal, Woolwich. Considerable experience with this method of projection showed that the fragments could be projected at the velocity used with an accuracy indicated by the following measurements made on successive shots: 1,474, 1,457, 1,468, 1,492, 1,453 m/sec. The mean value of these measurements (1,469 m/sec.) was in fact taken as the incident fragment velocity at the steel plates, as it was not convenient to measure this simultaneously with the velocities after perforation.

---

* Now with Messrs. Imperial Chemical Industries, Ltd., General Chemicals Division, Widnes, Lancs

The mild steel plates were all 15·25 cm. (6 in.) square, and were mounted for convenience in a stout wooden framework. The direction of attack was always normal to the plate surface, and, apart from a few cases which were easily distinguished, impact was made with the full square face of the fragment.

A Cooper inertialess chronograph (Cooper 1929), depending in principle on a number of small discharge lamps containing argon, was used for the velocity measurements. The momentary reduction of a resistance in series with any given lamp by an ' event ', in this case the passage of a fragment through a screen, causes a brief but intense flash. Images of the lamps are focused on photographic paper fixed to a rapidly rotating drum, intervals between events being deduced, as in other chronographs, by measurement of the distances between the marks on the record when the drum speed is known. Actually a time trace is produced on the record by the flashing of one of the lamps precisely every millisecond through a circuit controlled by an accurate invar steel tuning fork; thus errors due to changes in the dimensions of the photographic paper during processing are eliminated. The writing speed of the chronograph used in these experiments was about 25 mm/msec.

During its impact with the target plate the fragment was broken into a number of pieces, carrying with it pieces from the hole made in the plate. The main pieces of projectile and plate did not continue in quite the same direction, but were distributed within a cone of fairly small angle; also their velocities were not quite the same, as will be seen. Because of this, the usual method of timing intervals between successive arrivals at screens placed a known distance apart could not be used satisfactorily in its simplest form for the determination of the residual velocities. The principle of the method finally adopted was to record the impact with the target plate, and then the individual arrivals of the emerging pieces at one large screen a known distance away (2 or 3 feet). This screen was composed of strawboard sheets, forming a pack 2 ft. square and about 6 in. thick, its face being divided into six equal rectangular sections, 12 in. × 8 in. In each of these sections narrow tinfoil strips were arranged parallel to each other to form a uniform ' grid ', alternate strips being connected together at either end to make two interlocking but electrically independent systems. Thus each section was a separate recording screen which would be ' shorted ' momentarily by the passage of any piece of metal above a certain size. By spacing the strips so that very small pieces would not be recorded, and by connecting each of the six sections to different lamps of the chronograph, records were obtained which were fairly easy to analyse. Moreover, the pieces whose impact had been recorded could be recovered afterwards from the lower layers of the pack and identified.

The impact of the attacking fragment with the target plate was recorded by making the plate itself a recording screen. To do this, another grid of tinfoil strips was attached to a thin sheet of paper stuck to the front of the plate and connected to yet another lamp of the chronograph.

From the known distance between the plate and strawboard pack, and the time intervals elapsing since the original impact, the velocities of the pieces of fragment and plate could be deduced. Corrections to the velocities were made where necessary for obliquity of trajectory, but for the main pieces which were recorded this correction was never greater than 4%, and was often not more than 2%.

§ 3. GENERAL OBSERVATIONS

The fragment always pierced the plate by removing a plug therefrom, penetration taking place by punching out, rather than by a flow of material around the fragment. Details of all the pieces of fragment and plate recovered in each experiment have not been reproduced, but some indication of the degree of breakage suffered may be obtained from the masses of the principal pieces of fragment F and plate P given in the Table, column (3). That most of the breakage occurred during impact with the plate, and not during the subsequent penetration into the strawboard pack, was clear from the fact that the pieces were usually recovered from different parts of the pack.

It is possible that the break-up might have been influenced by any directional properties, or ' grain ', acquired during manufacture of the rods from which the fragments were cut. No experiments were made to investigate this point, but it is unlikely that the residual velocities, which are of main interest here, would be much affected, for example, by using fragments cut in directions other than perpendicular to the rod axis.

All the plates showed ' cratering ', the metal at the edges of the hole protruding beyond the general level on either face. This suggests that a limited amount of lateral displacement took place during penetration, at least on the side of entry; the front surfaces of the pieces of fragment also showed evidence of lateral displacement. Around the edges of the hole ' scabbing ' was also observed, particularly on the side of emergence.

For the plates thicker than about 0·5 cm. the area of the hole was about double that of the square face of the fragment itself, and was more round than square in shape. Reduction of the plate thickness below 0·5 cm., however, caused the hole to correspond more and more closely, both in area and shape, to the fragment face. A number of small particles of comparatively low velocity (not tabulated) were scattered by the impact through fairly large angles, and it is reasonable to suppose that many of these originated from near the edges of the hole in the plate, although some obviously came from the fragment itself. Pieces must have missed the pack altogether, but only occasionally was it possible for these to have been appreciably larger than any actually recovered (marked with asterisk in the Table).

Only in shots 2, 5 and 6 could the square face of the fragment have been inclined appreciably to the plate at impact. In these cases the holes made by the fragment were smaller, and the attacking fragment less broken by the impact, than usual; also the residual velocities of the fragments were somewhat higher than would have been expected had they hit squarely.

Three fragments were fired at plates 0·21 cm. thick with a velocity of only 900 m/sec. (residual velocities were not measured). The holes made tended to be a little smaller than for the usual velocity, and very little breakage of the fragment occurred. In contrast with the plates attacked at the higher velocity, these three plates were appreciably distorted in the neighbourhood of the hole, bulging outwards in the direction of penetration.

Velocities of Principal Pieces of Attacking Fragment and Plate after Impact

(1)	(2)	(3)	(4)	(5)
1	0·846	4·27 F*	536	559
		6·94 P*	597	
		5·85 P	518	
2†	0·851	11·10 F	674	557
		11·87 P	546	
3	0·787	5·43 F	557	599
		4·44 F	515	
		14·47 P	614	
4	0·643	9·43 F	744	706
		10·21 P	731	
5†	0·648	10·89 F	930	701
		1·46 P*		
6†	0·638	11·55 F	750	710
		7·18 P	744	
7	0·488	5·88 F	863	842
		7·27 P	868	
8	0·478	5·27 F	823	852
		6·50 P	875	
9	0·338	5·10 F	951	1000
		4·44 P		
10	0·312	4·16 F	847	1029
		3·73 P*	1155	
11	0·206	9·66 F	1176	1161
		2·51 P	1184	
12	0·208	9·14 F	1204	1158
		2·54 P		

Incident fragment velocity ~ 1,469 m/sec.

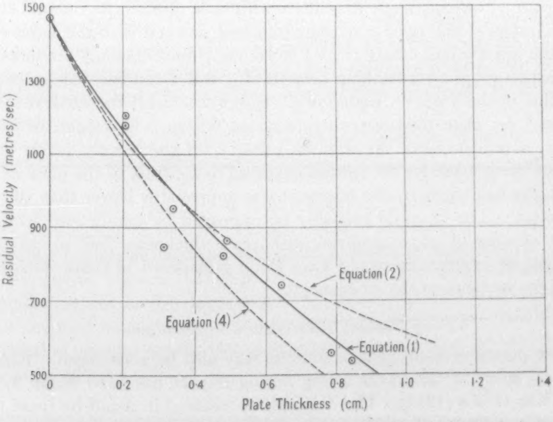
(1) Shot number; (2) Plate thickness (cm.); (3) Masses of principal pieces recovered (gm.); (4) Observed velocity (m/sec.); (5) Calculated velocity (m/sec.) (equation (1)).
F = Fragment. P = Plate.

* Only in these cases could there have been pieces appreciably larger than any of those recovered.

† Fragment did not hit squarely.

§ 4. RESIDUAL VELOCITIES

The velocities of the principal pieces of fragment and plate emerging after penetration are recorded in the Table. The residual fragment velocity, taken as that of the largest piece of attacking fragment for which a record was obtained, is plotted against plate thickness in the Figure; the three cases in which the fragment did not hit squarely are omitted.



Variation of residual fragment velocity with plate thickness.

It will be seen that the velocities of the principal pieces of plate carried away usually agreed fairly closely, considering the nature of the measurements, with the residual fragment velocity. The outstanding exception to this is shot 10, in which the velocity of the principal piece of plate removed was exceptionally high and the residual fragment velocity exceptionally low. It does not seem profitable to speculate upon the reason for this isolated result, however, and all the other residual velocity measurements are reasonably consistent with the fitted curve drawn in the Figure.

§ 5. DISCUSSION

Although the plates were only 6 in. square, it should be legitimate to regard the results as applying to plates of ' infinite ' extent, as any disturbance would not be reflected from the edges before penetration was complete.

The curve drawn in the Figure to fit the experimental results shows a slight but significant deviation from linearity, and is actually represented by the empirical equation

$$V_R = V_0 \exp(-1{\cdot}142x), \qquad \ldots\ldots(1)$$

where V_R is the residual velocity corresponding to a striking velocity V_0(1,469 m/sec.), and x is the plate thickness in centimetres. The actual values of V_R calculated from this equation are given in the last column of the Table. Obviously such an equation can only be used with incident velocities well above the critical for perforation, as in the present case, since there is no implication of a stopping thickness.

(i) *Elementary Interpretation*

The simplest attempt at a theoretical approach to this problem is made by considering the conservation of momentum before and after impact. The equation which can be derived by assuming momentum to be conserved between the attacking fragment, thickness t, and a piece of plate of area equal to that presented by the fragment on (square face) impact is

$$V_R = V_0 t/(t+x), \qquad \ldots\ldots(2)$$

if it be assumed also, in rough accordance with most of the experimental evidence, that the fragment and portion of plate removed proceed with the same velocity. When the appropriate values ($V_0 = 1{,}469$ m/sec., $t = 0{\cdot}711$ cm.) are substituted, this equation gives a relationship between V_R and x represented by the upper dotted line in the Figure. Equation (2) agrees well with the observed residual velocities* for plate thicknesses up to about 0·5 cm. As might be expected, however, it overestimates the residual velocity for the thicker plates used, no allowance being made for the motion imparted to the bulk of the plate or for the fact that the hole made by the fragment was appreciably larger than the area of presentation. The fit could probably be improved by taking into account the increase of area of plate removed with plate thickness, but to do this the mechanism of penetration would have to be considered in more detail than is justified by the scope of the experiments.

(ii) *General Theoretical Considerations*

A few points of more general interest may also be considered. Recently a theoretical study of the forces acting during impact has been made by White and Le Van Griffis (1948). The only case considered in detail by these authors is that of a cylinder of yielding material impinging axially on an unyielding plane surface. In the limit, for interaction between solid bodies at very high relative velocities, they assume in common with the authors of other recent papers, e.g. Grumminger (1948), and Birkhoff *et al.* (1948), that the forces may be obtained from hydrodynamical considerations. Thus for a striking velocity V greater than the maximum wave velocity in the yielding substance they take the force F as being identical with that exerted by a fluid, i.e.

---

* Strictly the velocity components taken normal to the plate should be used for this comparison, rather than the actual velocities (corrected for obliquity), but the difference is too small to be important, see § 2.

$$F = \rho A V^2, \qquad \ldots\ldots(3)$$

where A is the area of cross section of the cylinder and ρ the density of its material. Actually it is shown that flow can occur at impact velocities considerably lower than the maximum wave velocity. In appearance the type of high velocity impact visualized by White and Le Van Griffis is similar to that shown in the high-speed photographs obtained by Schardin and Struth (1937) of the impact of a lead-filled bullet of velocity 1,000 m/sec. against strong armour plate. The photographs show the lead flowing outwards along the plate surface, but, as the authors point out, the lead should be partly vaporized in this case. Finally it is mentioned by White (1943) that when both target and projectile yield, the effective velocity is reduced, being only half the actual velocity of impact in the special case when the two are of the same material.

Although in the present experiments there was evidence of some flow in both the plate and the front surface of the fragment, the mechanism of penetration was mainly by punching out, or plug formation, rather than by pushing aside of the plate material. Any attempt at precise calculation of the motion of the plug from rest would probably involve an extension of the methods used by Taylor (1948) to calculate the retardation of a flat-ended projectile during impact with a target. However, removal of the portion of plate ahead of the impinging fragment should begin after an interval of order given by the time taken for a disturbance of velocity about 6,000 m/sec. (see, for example, Pack, Evans and James 1948) to pass through the plate. This suggests that a fragment of velocity 1,469 m/sec. might have little opportunity for 'flow' penetration under the present conditions of experiment. Nevertheless, plug formation can, of course, only occur if the stress conditions around the portion of plate in the vicinity of the fragment are suitable, whatever the relative velocities of impact and disturbance.

The lower dotted line in the Figure is obtained from the equation

$$V_R = V_0 \exp(-\rho A x / m), \qquad \ldots\ldots(4)$$

calculated for a 'hard' fragment of mass m and area of presentation A penetrating a yielding substance of density ρ which offers a resistance given, in the ideal case for a flat contact surface and high Mach number, by equation (3). Actually the discrepancies between the experimental values and those calculated from equation (4) are not so very large in spite of the obvious differences between the real and assumed mechanisms of penetration, but a closer approximation should be provided in practice for the case of penetration of lead sheet by steel fragments of very high velocity.

ACKNOWLEDGMENTS

Grateful acknowledgment is due to the Chief Scientist, Ministry of Supply, for permission to publish this paper, and to Dr. R. Ferguson, Mr. A. Warren and other colleagues for the benefit of discussion.

The experimental work was carried out at the Safety in Mines Research Station, Buxton, and the author is indebted to the Director and members of the staff for the facilities which were so generously placed at his disposal. Also it is desired to thank Messrs. R. Torry, G. Ashmore, V. O. Hardy, and R. B. Priest for much assistance with the field work.

REFERENCES

Bashforth, F., 1873, *Motion of Projectiles* (London : Asher and Co.,) Chapter 6.

Bauer, W., 1926, *Ann. Phys., Lpz.*, **80**, 232.

Birkhoff, G., MacDougall, D. P., Pugh, E. M., and Taylor, Sir Geoffrey, 1948, *J. Appl. Phys.*, **19**, 563.

Cooper, P. A., 1929, *Phil. Mag.*, **8**, 1100.

Gildmeister, M., and Strehl, H., 1905, *Ann. Phys., Lpz.*, **18**, 567.

Grumminger, G., 1948, *J. Appl. Phys.*, **19**, 947.

Honda, K., Takermaé, G., and Watanabe, T., 1930, *Sci. Rep. Tohoku Imp. Univ.*, **19**, 703.

Pack, D. C., Evans, W. M., and James, H. J., 1948, *Proc. Phys. Soc.*, **60**, 1.

Resal, H., 1895, *C. R. Acad. Sci., Paris*, **120**, 397.

Schardin, H., and Struth, W., 1937, *Z. tech. Phys.*, **18**, 474.

Taylor, Sir Geoffrey, 1948, *Proc. Roy. Soc.* **A, 194**, 289.

White, M. P., 1943, *N.D.R.C. Report* A, 157.

White, M. P., and Le Van Griffis, 1948, *J. Appl. Mech.*, **15**, 256.

An Air-Cored Synchro-Cyclotron for 400 Mev. Protons

By L. RIDDIFORD

Department of Physics, The University, Birmingham

Communicated by M. L. Oliphant; MS. received 12th October 1950

ABSTRACT. It is considered that the most reasonable proposition for accelerating protons to very high energies, using an instrument which employs a magnetic field produced by an air-cored coil, is to operate the machine as a synchro-cyclotron. An investigation of the energies available from such accelerators for given amounts of stored magnetic-field energy is carried out on the basis that the limit is set by the mechanical forces exerted on the coil. It is shown that the power requirements are comparable with those of large iron-cored accelerators at present under construction. In a machine with a useful orbit of radius 11·8 cm., it is possible to provide a field of $2{\cdot}7 \times 10^5$ gauss over the circle enclosed by that orbit, so that protons would be accelerated to an energy of 400 Mev.

§ 1. INTRODUCTION

The energy obtainable from any accelerator which employs a magnetic field is limited only by the value of the product BR, presuming all other technical difficulties can be overcome. Until now the general practice has been to use iron-cored magnets, which sets an upper limit of about 15,000 gauss to the value of B, and then to make R as large as the financial resources available will permit. So far as the author is aware, the only air-cored machine in operation anywhere in the world at the moment is the G.E.C. synchrotron (Jones *et al.* 1950), running at low energy as a betatron; it should, however, be mentioned that Hill (1947) has succeeded in accelerating electrons to 100 kv. with a fixed-frequency air-cored electron cyclotron (or magnetic resonance accelerator) in which the relativistic mass increase is taken care of by using an increasing magnetic field (superimposed on a static field) instead of frequency modulation. But it seems clear that many such accelerators will be built in the future, for by the use of air-cored magnets it is possible to reduce the physical size of the accelerating device to very modest dimensions indeed. Of course, large comparable quantities of electrical energy are required in either case. Two attractive features of the air-cored machine are the substantial reduction in cost on all items except the source of power, particularly in steel, building space and engineering effort, and the desirability in an experimental physics laboratory of having the tools of research in as convenient a form as possible.

The technical reasons for neglecting an air-cored accelerator are many, among the difficulties being: (i) the high precision with which the conductors must be placed in position to give the correct field distribution, (ii) the magnitude of the forces on these conductors which require them to be very firmly held, and in fact in the limit approach the yield strength of the material itself, (iii) temperature rise in the conductor due to large quantities of power being delivered into a small volume, (iv) switching problems, and leads and connections between turns, (v) injection and extraction difficulties, (vi) radio-frequency problems—in the case of electron accelerators these do not matter so much, but for heavier particles accelerators of the synchrotron type would require precise frequency changes over a large range in the 100 Mc/s. region.

The energy of an electron is in the extreme relativistic region when it amounts to is only a few million electron volts, but for protons it is necessary to go to 3,000 MeV. before this is so. Further, it is possible to accelerate electrons to such energies by betatron action. So an air-cored electron synchrotron may be a practical possibility, and in fact Blewett (1947 a, b) published the results of theoretical and experimental work on such an accelerator, carried out at the Research Laboratory of the General Electric Company. The energy aimed for at that time was 100 MeV., with a field of 20,000 gauss at an orbit of radius 7 inches. A more recent internal report from that Laboratory suggests that it is hoped to raise this to 1,000 MeV., which would require a field of 2×10^5 gauss for the same radius. At such energies and radii, however, the loss of electron energy by radiation is very serious indeed (Schiff 1946), but radiation losses are negligible in proton accelerators.

An alternative to initial betatron acceleration may lie in the harmonic air-cored electron synchrotron proposed by Kaiser and Tuck (1948). Lawson (1949) has discussed the possibility of the application of such harmonic acceleration to protons, but concludes that its practicability is doubtful.

When these considerations are taken into account, it appears that for a pure air-cored accelerator the synchro-cyclotron is the most reasonable proposition for heavy particles, although of course somewhat less efficient than an air-cored synchrotron of the same size would be from the point of view of stored energy.

§ 2. COIL DESIGN

The form of the coil is defined by the nature of the mechanical forces involved, and ease of winding. A simple cylindrical coil is the easiest to wind. Further, since clear space for a vacuum box in which to accelerate the particles is required, no possible means of resisting the tremendous forces between conductors placed above and below this space is available. Consequently, it is necessary to make the coil in the form of a hollow cylinder. We proceed to define the optimum geometry of such a coil.

Calculations by Cockcroft (1928) have yielded a series of curves enabling one to specify the best coil shape for a maximum central field $H(0,0)$ using a given amount of electrical power. The important approximate result which may be deduced from these curves is that for a coil of the type shown in Figure 1 the best geometry is

$$z_2 = 0{\cdot}63x_2. \qquad \ldots\ldots(1\,a)$$

On this basis, it is necessary to calculate the maximum value of $H(0,0)$ obtainable from coils of various volume and internal diameter, when the limit is set by the fact that the stresses set up in the coil should at no point exceed the yield strength of the conducting material. This means that for a given HR (or particle energy), the stored magnetic energy is as small as possible; for in any magnetic field accelerator this stored energy increases linearly with R, for a given HR, if the shape of the vacuum-box remains fixed, but its size increases in proportion to R.

Clearly the stresses are greatest at the point C $(x_2, 0)$ in Figure 1. The largest forces are circumferential, but these do not enter into the discussion as the coil can be prevented from bursting by holding it in a stout retaining ring. Kapitza (1927) has shown how to calculate the radial and axial stresses at any point (x, z) in the coil, and Cockcroft (1928) has plotted curves calculated on this basis.

The most suitable conducting material available is cadmium–copper, whose strength may be twice that of copper for a reduction of only 15% in electrical conductivity. Beryllium–copper has a conductivity only 25% of pure copper. The greater of the x and z components of stress has in all cases been set equal to 4,000 kg. cm^{-2} (25·4 tons inch^{-2}). Values of $H(0,0)$ follow from the allowed maximum values of $10\,NI$, the ampere turns. Since for all coils under consideration it can be assumed that the field $H(x_1, 0)$ is also approximately equal to $H(0,0)$, the maximum value of HR is about $H(0,0)x_1$.

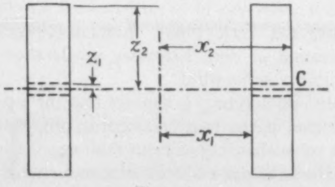


Figure 1.

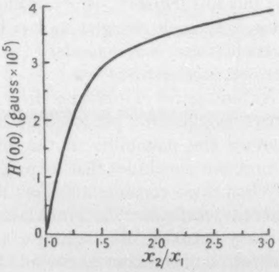


Figure 2.

Figure 2 is a plot of $H(0,0)$ against x_2/x_1 which holds for coils of any volume, the maximum value of $H(0,0)$ depending only on coil shape. Of importance is the knee in the curve at $x_2/x_1 = 1{\cdot}45$, which indicates that it is inefficient to increase the volume of the coil beyond that corresponding to $x_2/x_1 = 1{\cdot}45$ unless compelled to do so by temperature rise considerations. It is physically possible that at the limit set by the forces on the conductors there is an optimum type of coil giving a maximum HR (or Hx_1) for a given amount of stored energy W. W, which determines the peak ampere-turns required, depends on both x_1 and x_2/x_1; it is thus of interest to know whether it is best to keep x_1 small, and use a coil of large volume to raise $H(0,0)$, or to use a larger x_1 and a coil of smaller volume. If we plot a series of graphs of W against x_2/x_1 for various x_1, and from these and Figure 2 plot values of Hx_1 as a function of x_2/x_1 for a series of values of W, it is found that in the latter figure fairly flat maxima do occur, all at x_2/x_1 equal to about 1·45, as expected from previous considerations.

Consequently, once x_1 has been chosen, the coil shape is completely defined by the two simple equations

$$x_2/x_1 = 1{\cdot}45, \quad z_2/x_2 = 0{\cdot}63. \qquad \ldots\ldots(1\,b)$$

Figure 3 is a plot of $(HR)_{\max}$ against W for the air-cored synchro-cyclotron, and a proton synchrotron curve based on the Birmingham machine is also shown, assuming the latter produces protons of energy 1,000 MeV. It should be remembered that the particles can only be accelerated out to a radius somewhat less than x_1.

To prevent the coil bursting it must be enclosed in a retaining ring. The ring is subject to a uniform radial stress P_x of 4,000 kg. cm^{-2}. The stresses in it are calculable and the result is that the retaining ring must be about as large as the coil itself. The retaining ring should be bound tightly from strip to reduce eddy current losses and circulating currents from induced voltages of order 100 volts per turn. Similar care is to be taken with other parts of the assembly. The coil winding itself should not exceed about 1 cm. square section because of skin effects. Proximity effects will no doubt increase power losses anyway.

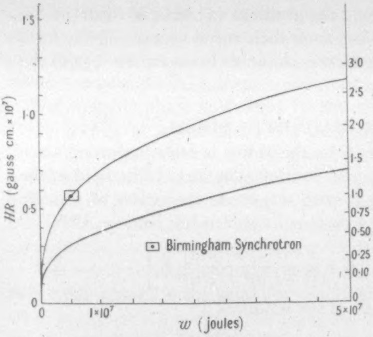


Figure 3.

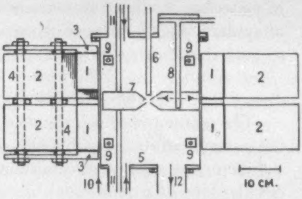


Figure 4.

1. Cd–Cu coil
2. Retaining ring
3. Clamps
4. Bolts
5. Vacuum box
6. Ion source
7. Dee
8. Probe
9. Correcting coil
10. Cooling
11. R.F. power in
12. To pumps

§ 3. SOURCE OF ELECTRICAL ENERGY

The best way of storing the energy required to pulse a large magnetic field lies in the rotational inertia of a large electrical generator (supplemented by flywheel energy if necessary). An alternator appears more suitable than a D.C. generator since in the latter the available current is limited greatly by the presence of the commutator. Further, alternators are cheaper, of smaller physical size, and permit of ease in switching off the current after the accelerating cycle, as the field falls to zero of its own accord. Also the pulse length is independent of the time-constant of the coil.

Consider a coil of dimensions: $x_1=12{\cdot}5$ cm., $x_2=1{\cdot}45x_1=18{\cdot}1$ cm., $z_2=0{\cdot}63x_2=11{\cdot}4$ cm., $z_1=0{\cdot}06x_1=0{\cdot}75$ cm. Aiming at 400 MEV. protons, we require $HR=3{\cdot}2\times10^6$ gauss cm., which means $H=2{\cdot}71\times10^5$ gauss for a useful $R=11{\cdot}8$ cm., and consequently $10\,NI=8{\cdot}60\times10^6$ ampere-turns. For this coil, $L_0=0{\cdot}21\times10^{-6}N^2$ henry, and $R=1{\cdot}93\times10^{-6}N^2$ ohm, taking the resistivity of cadmium–copper as $2{\cdot}4\times10^{-6}$ ohm cm. The stored energy $W=7{\cdot}8\times10^6$ joules ($1{\cdot}86\times10^6$ calories) and the time constant L_0/R of the coil is 0·108 second. Its volume is 12·3 litres and mass of cadmium–copper 241 lb. The coil is drawn to scale in Figure 4, which is a schematic diagram of the proposed accelerator.

For a 25-c/s. alternator, and $I=10^5$ amp., 86 turns are required and the peak voltage has to be 12,300. ←

Heating must be small so as to maintain good insulation and keep the resistance of the coil down. Good insulation would stand 160° C., but it would be unwise to take the temperature above the much lower value at which tertiary slip in the metal occurs. For the coil considered, the temperature rise during half a sine wave of the 25-c/s. alternator is $\theta=35°$ C. Clearly then the temperature rise is not serious, but the coil must, of course, be cooled to keep the duty cycle up. The logical way to do this is to use the conductor itself as a water pipe. However, this is not possible as the enormous circumferential stresses would in time force the copper into the void. One way around the problem would be to form the coil in pancakes, insulated from each other and from their metal spacers, in the form of spiders, placed between them. The narrow channels between the legs of the spiders would carry the cooling fluid.

§ 4. SHAPE OF THE MAGNETIC FIELD

The manner in which the field depends on the radius is most important since the particle orbits are stable only if the field is a focusing one. The field in the z-direction at a point of coordinates (r, z) with respect to the centre of a single circular loop of wire of radius a, carrying current I E.M.U. is (cf. Smythe 1939)

$$H_z(r,z)=\frac{2I}{a[(1+r/a)^2+(z/a)^2]^{1/2}}\left[K+\frac{1-(r/a)^2-(z/a)^2}{(1-r/a)^2+(z/a)^2}E\right], \quad \ldots\ldots(2)$$

where K, E are complete elliptic integrals of the modulus

$$k^2=4(r/a)[(1+r/a)^2+(z/a)^2]^{-1}.$$

Precise calculations for the field of a cylindrical coil of the type concerned here would only be possible in the form of a converging series. For the present one has been content to integrate equation (2) graphically over both radial and axial coordinates, which gives about $\frac{1}{2}\%$ accuracy if done carefully with a planimeter. We require

$$H_z(r,z)=4i\int_{(a/r)_1}^{(a/r)_2}\int_{(z/a)_1}^{(z/a)_2}\frac{d(z/a)\,d(a/r)}{[(1+r/a)^2+(z/a)^2]^{1/2}}\left[K+\frac{1-(r/a)^2-(z/a)^2}{(1-r/a)^2+(z/a)^2}E\right], \quad \ldots\ldots(3)$$

where the z/a integration must be carried out first; i is the current density in E.M.U.

The coil has to be wound in two halves, one above and the other below the median plane, firstly to allow a beam of energetic particles to emerge without obstruction, and secondly because the field of a coil without a central gap is strongly defocusing. The result of an investigation of equation (3) is that the field is always a defocusing one over the central region of the useful circle, but that it may be either focusing or defocusing in the outer region, depending rather sensitively on the separation of the two halves of the coil. To make best use of the coil this dimension should be optimized.

In any event, it is necessary to correct the central defocusing by an auxiliary coil. The magnetic energy required for this field is of order only $0{\cdot}01\%$ of that in the main pulsed field. The attainment and measurement of a correctly shaped field presents a most difficult experimental problem, and would require careful work with both electrolytic tank and nuclear resonance apparatus. Frequencies of about 1,000 Mc/s. are required for proton resonances.

Plots of the function under the integral sign in equation (3), as a function of r/a, z/a, for fixed values of z/a, a/r, are particularly useful for they enable one to see quickly in what manner it is possible to change the field shape in a particular way. Further, they indicate immediately how to place correcting coils to compensate for defocusing or asymmetries. According to Henrich, Sewell and Vale (1949), the beam intensity in the 184-in. cyclotron may be increased by a factor of ten by shaping the central region of field.

§ 5. RADIO FREQUENCY

The limited frequency range required, and wide tolerances, are fundamental advantages of this machine over an air-cored proton synchrotron. For the field of $2{\cdot}7 \times 10^5$ gauss involved here, the initial frequency is 417 Mc/s., and at 400 MEV. where $\beta = 0{\cdot}72$, it must have fallen to 290 Mc/s., a change of only 30%.

For a stable phase ϕ_s defined by the condition $\sin\phi_s = 0{\cdot}5$, due to Bohm and Foldy (1947), at which there is a maximum acceptance of particles, and peak dee-to-dee voltage 10 kv., 8×10^4 particle revolutions are needed for acceleration to full energy, so that the acceleration time is about 0·0002 second. Consequently, it should be possible to accelerate more than one bunch of particles during each field pulse, particularly if the top of the pulse were flattened, as in Kapitza's experiments, by rearranging the alternator windings.

It is worth noting that the frequency range over which particles can be accepted into stable phases is the same for the cyclotron under discussion as it is for any other proton synchro-cyclotron using the same dee voltage. Putting numerical values into Bohm and Foldy's equation, we have $\pm\Delta\omega_s/\omega_s \simeq 1{\cdot}91 \times 10^{-4} V^{1/2}$, where V is the peak dee potential in E.S.U. For $V = 10$ kv., $\Delta\omega_s/\omega_s = \pm 1{\cdot}1 \times 10^{-3} = 0{\cdot}22\%$.

It may well be better to avoid the need for frequency modulation and to use a fixed frequency in the manner of Hill (1947). For protons reaching 400 MEV. when the magnetic field is at its maximum value of $2{\cdot}7 \times 10^5$ gauss, it is required to add an average of 215 volts in each of $1{\cdot}95 \times 10^6$ revolutions. The particles are captured into phase-stable orbits when the field is $1{\cdot}88 \times 10^5$ gauss, and then accelerated by a radio-frequency field of 390 Mc/s. Under these conditions the rise-time of the field-pulse is sufficiently slow for synchronous acceleration (Veksler 1945, McMillan 1945). The choice then, seems to lie between the comfort of fixed-frequency operation and the advantage of the production of multiple pulses, associated with a frequency-modulated arrangement. In fact, it should be quite possible to employ the two methods simultaneously, thus increasing the efficiency of the instrument. The matter is open to experiment.

ACKNOWLEDGMENTS

The author wishes to thank Professor M. L. Oliphant for his encouragement and advice, and Mr. J. W. Blamey for several discussions. It is a pleasure to acknowledge that he has held an Exhibition of 1851 Science Scholarship, and a Dafydd Lewis Science Scholarship, during the course of time in which this work was done.

REFERENCES

BLEWETT, J. P., 1947 a, *J. Appl. Phys.*, **18,** 968; 1947 b, *Ibid.*, **18,** 976.
BOHM, D., and FOLDY, L. L., 1947, *Phys. Rev.*, **72,** 649.
COCKCROFT, J. D., 1928, *Phil. Trans. Roy. Soc.* A, **227,** 317.
HENRICH, L. R., SEWELL, D. C., and VALE, J., 1949, *Rev. Sci. Instrum.*, **20,** 887.
HILL, R. D., 1947, *Nature, Lond.*, **159,** 774.
JONES, W. B., *et al.*, 1950, *Phys. Rev.*, **78,** 60.
KAISER, T. R., and TUCK, J. L., 1948, *Nature, Lond.*, **162,** 616.
KAPITZA, P., 1927, *Proc. Roy. Soc.* A, **115,** 658.
LAWSON, J. D., 1949, Atomic Energy Research Establishment Memo G/M119.
McMILLAN, E. M., 1945, *Phys. Rev.*, **48,** 143.
SCHIFF, L. I., 1946, *Rev. Sci. Instrum.*, **17,** 12.
SMYTHE, W. R., 1939, *Static and Dynamic Electricity* (New York : McGraw-Hill), p. 266.
VEKSLER, V., 1945, *J. Phys., U.S.S.R.*, **9,** 153.

Interferometric Studies on the Vibration of Piezoelectric Plates

BY S. TOLANSKY AND W. BARDSLEY
Royal Holloway College (University of London), Egham, Surrey

MS. received 9th February 1950; *in amended form 7th November* 1950

ABSTRACT. A description is given of the application of a number of multiple-beam interference methods to the study of the piezoelectric oscillations of some quartz crystals. The distribution of surface nodes and anti-nodes of motion and the amplitudes of oscillation are revealed by simple methods. By employing a stroboscopic illumination technique the phase relations between the various vibrating regions of the crystal suface are revealed. The nodal regions are shown to be at rest to within less than 50 A. and the surface deformations are revealed to at least a similar order of accuracy. By a modified technique fringes are obtained through the crystal body and these reveal changes in the optical thickness. One crystal developed sharp surface discontinuities after being set into oscillation, indicating that crystallographic glide or slip had taken place. This is shown to be closely related to the familiar shattering of quartz crystals at excessive amplitudes of oscillation.

§ 1. INTRODUCTION

In this paper a preliminary description is given of the application of multiple-beam interference methods to the study of the vibrations of piezoelectric quartz crystal plates. These methods are superior to those used by earlier workers.

In the formidable literature on the theory and practice of piezoelectric quartz crystal oscillators the only significant interference experiments on the nature of the mechanical oscillations of the crystal are those reported by Osterburg (1929 *et seq.*) and Dye (1932), and both investigators used only two-beam interference methods.

Osterburg (1929, 1932) set up a Michelson interferometer with an oscillating quartz crystal plate as one of the end mirrors, and in succeeding papers devised a compound interferometer whereby simultaneous observations of opposite faces (1933 a) and also of adjacent faces (1933 b) could be carried out. This was then followed by a paper (1934) in which the crystal plate was placed in one of the light beams of a Michelson interferometer and a direct determination made of the resultant change of the optical path arising when the plate was oscillating.

The method used by Dye (1932) closely resembles *one* of the techniques we employ, but with the restrictions imposed by two-beam interferometry. Dye used localized reflection Fizeau fringes formed between the lower surface of a quartz crystal plate and a glass flat on which the crystal rested. He developed a stroboscopic technique whereby the fringes were produced by a light source which was illuminated periodically at the same frequency as the crystal.

§ 2. EXPERIMENTAL

The quartz plates used in these studies have mostly been of square or rectangular parallelepiped form with a thickness small in comparison with the other two dimensions. The thinner plates could not be polished flat, and showed curvatures of the order of one or two light waves.

The face to be studied was coated with the usual type of high quality silvering employed for multiple-beam interferometry (Tolansky 1948) and then placed gently on a similarly coated glass flat.

In the initial experiments the back surface of the crystal was lightly silvered to act as an electrode, light spring contact being made to this silver film, contact to the silvered surface of the glass flat providing the other electrode. Strip electrodes of various forms were then tried but it was finally found satisfactory in the majority of cases to use for one electrode a wide mesh grid of a few quite thin parallel wires some millimetres above the back face of the crystal, whilst retaining the silver coating on the glass flat as the other electrode. Crystal plates were excited either with Hartley or Pierce type oscillators.

Experiments of two kinds have been carried out. In the one the fringes are viewed with a steady arc source, and in the second a stroboscopic arrangement has been developed in which a source has intensity fluctuations imposed on it with the same frequency as that of the oscillating crystal.

In the stroboscope experiments the light source used was driven from the oscillator. A Geissler tube was connected in series with a D.C. polarizing voltage and a parallel resonance tuned circuit coupled inductively to the Hartley tank circuit. By adjusting the polarizing voltage it was possible to flash the light source with the same frequency as that at which the quartz plate was being driven. Phase shifts between the two circuits could be introduced enabling the light pulse to occur at various positions within the amplitude range. A suitable discharge can be obtained either with helium gas or with a mercury–argon gas mixture. Stroboscopic illumination up to 1,000,000 c/s. was obtainable.

§ 3. OBSERVATIONS ON SURFACE VIBRATIONS

Through the courtesy of Dr. Radley of the G.P.O. Research Station, Dollis Hill, we have available a number of quartz plates of various crystallographic cuts. Further, through the courtesy of Dr. Essen of the National Physical Laboratory we have some of the well-worked quartz plates used by the late Dr. Dye in his original pioneer work (Dr. Essen originally prepared the surfaces). Many of these crystals have now been examined interferometrically, and from these data some of the more significant features of general interest are presented here.

A more detailed analysis of the studies made on the various motions of specifically selected widely used crystal cuts is reserved for a succeeding communication. Several photographs are presented since they reveal the phenomena in a very clear manner. Except where otherwise indicated, the interferograms are multiple-beam Fizeau fringes taken with $\lambda 5461$ (Hg).

Plate I (*a*)* shows the fringes given by the surface of a rectangular quartz plate at rest. The surface has been figured to a curvature to give semi-circular fringes. Plate I (*b*) shows the appearance when the crystal is set into oscillation with a small amplitude. It will be seen that in those regions which are moving, because of the change of air-film thickness between crystal and flat, the fringes broaden out. At the nodal regions the fringes remain narrow. Thus the interferogram reveals the local distribution of vibration, and precise amplitude values are readily measured. The movement revealed is only the resolved component perpendicular to the surface, and the technique, like most others, does not reveal motions parallel to the surface.

Plate I (*c*) shows fringes given by the same crystal, but vibrating in a distinctly different mode with a number of clearly distinguishable nodal regions. With such widely spaced fringes the true nodal regions may be easily missed, in so far as it is then purely an accident of the particular angle of inclination between flat and crystal which determines whether or not a fringe passes through a true nodal region. If the amplitude is made too large as in Plate I (*d*) the fringes overlap and eventually visibility suffers so that interpretation becomes impossible.

Plate I (*e*) is an example of 'crossed fringes' obtained as follows: The flat-crystal inclination is set with fringes running in the vertical direction and the vibration fringes are photographed. The wedge angle between crystal and flat is then altered to bring the fringes running horizontally and once more the crystal is set oscillating with the same frequency. Without moving the camera or the photographic plate, a second exposure is superposed on the first, the result being the pattern of Plate I (*e*). A picture such as this indicates that on a single 'high dispersion' photograph the true node is rarely on a fringe.

Any ambiguities about node location are eliminated when closely-spaced fringes are used and several examples will be illustrated under such conditions, for instance, as in Plate I (*f*), in which the nodel and vibrating regions are clearly and distinctly shown. (This was an early photograph taken with a brass bar as upper electrode above the back surface, visible as a black strip.)

The amplitudes in any particular region of any crystal can be measured with precision and it is estimated that in most cases the nodal regions are at rest to within certainly less than 50 A., this limiting value being set by the fringe width.

With moderate amplitudes there is practically no measurable displacement of the crystal as a whole relative to the position occupied when at rest (Plate II (*a*)). In this, two sets of fringes are superposed and photographed as follows: first the fringes when at rest were photographed, then by a second exposure, without altering either crystal or camera, the fringes of the vibrating system were superposed. There is a narrow 'rest' fringe passing almost exactly symmetrically down the centre of each 'vibration' fringe. Any displacement of the 'centre of gravity' of the broadened fringes from the original narrow 'rest' fringes is hardly measurable, and since fringe width is about one-fiftieth of an order, i.e. 50 A., it follows that the vibrating crystal as a whole has maintained its position relative to the flat to within less than 50 A. With larger amplitudes bending of the crystal is first noted and then bodily movement appears.

---

* For all Plates see back of issue.

With certain crystals, examination with many fringes in the field reveals in an elegant manner a multiplicity of nodes and establishes the upper limit to the amplitudes. Examples are shown in Plates II (*b*) and II (*c*), the former being a rectangular crystal of GT cut and the latter a rectangular crystal of X cut. The frequency in the former case, a harmonic of the fundamental, approached 3 Mc/s. The uniformity of amplitude and pure regular node distribution are well shown. The second crystal according to calculation should be exhibiting a harmonic of the longitudinal compression mode in the length direction. The figuring at the edges is not perfect and this accounts for the distortions in the pattern.

It will now be considered whether the velocities of the oscillating faces contribute to the observed fringe widths by introducing Doppler effects. Taking as an extreme example a frequency of 100 kc/s. and a maximum amplitude of some two orders, say 5,000 A., then the calculated velocity of the plate surface has the surprisingly low value of only 16 cm/sec. Clearly therefore Doppler effects are entirely ruled out. In any case the chromatic resolving power of the interferometer is very low, furthermore any possible refractive index changes in the air film can have no visible effect on the observations.

It is certain that the interferometric contours shown are indeed the true crystal surface contours during oscillation.

§ 4. REFLECTION FRINGES

It was considered desirable to make experiments on oscillating crystals which were cemented to metal plates, hence observations were conducted to see if reflection methods could be adopted. Some difficulty was anticipated in connection with fringe visibility for in the reflected system the fringes for the quartz plate at rest consist of fine *dark* lines on a bright background. When the plate oscillates it is clear that as the fringes move into the bright background this will 'dilute' the visibility considerably. In spite of this effect, by judicious choice of silverings we have succeeded in using multiple-beam reflected fringes to study oscillations. Plate II (*d*) shows such reflection fringes at rest and Plate II (*e*) the appearance when the quartz plate is vibrating. The precision of measurement of these fringes is not quite equal to that of the transmitted system but approaches it. (The low visibility 'ghost' fringes, due to the wedge angle between the two surfaces of the optical flat, do not introduce difficulties.)

§ 5. STROBOSCOPE STUDIES

Dye (1932) found in his experiments that with two-beam interferometry the sinusoidal intensity distribution of the fringes restricted most observations. Whilst these fringes did reveal major nodal regions of the vibrating surface it was difficult to arrive at any precise picture of the surface motion at any general point and the relative phases of the instantaneous amplitudes at any point were indeterminate. Dye therefore developed a stroboscope device whereby a fraction of the oscillatory power was employed in driving a discharge tube which was then used as the light source for production of the fringes. We have applied this technique which when combined with multiple-beams gives additional important information.

The information revealed by the stroboscopic multiple-beam approach is complex and detailed interpretation is formidable. Two examples have been selected to illustrate the information made available. Plate II (*f*) shows the stroboscope picture of a moving surface. It will be recognized that the stroboscopic fringes in effect arrest the movement at the selected position within the whole oscillation and the important feature is that the relative phase directions of the instantaneous motion of the different local areas are revealed. With a stationary source the broadened fringes reveal the regions of oscillation, but one has no possible means of determining in which directions the different areas are displaced. In the stroboscope picture one sees quite clearly the instantaneous directions. The relative amplitudes, i.e. the deviations from the rest fringe position, are arbitrarily determined by the particular phase selected within the complete oscillation.

The effect of varying the stroboscope phase to pick out a different section of the total oscillation is shown by Plates II (*g*) and II (*h*); they show two near but different phase positions. There are notable fringe changes in going from (*g*) to (*h*). Plate II (*h*) is furthermore a ' crossed ' fringe picture, for superposed here are two distinct sets of fringes. These consist of stroboscopic vibration fringes at a phase position close to that of Plate II (*g*) and on these are superposed the smooth curved fringes shown by the crystal when at rest. (The original surface was crudely worked and was not flat.) The combination photograph reveals in a striking way the instantaneous behaviour of this particular vibration at the selected phase and shows too that some of the ' nodes ' are effectively at rest whilst others have a residual small secondary amplitude. An absolute node is only indicated at the points where the two systems touch each other tangentially.

§6. VIBRATIONS WITHIN CRYSTALS

The discussion has been confined so far to studies of the *surface* movements of the oscillating quartz plates. In addition exploratory experiments have been made on the examination of the vibrations within the body of the crystal. Two methods offer themselves. In the first both major surfaces of the crystal are polished reasonably flat and nearly parallel, and then silvered. Fringes are produced with the path difference within the crystal plate between the two surfaces. In the second method fringes are separately formed on each surface by matching each against a separate flat, and the fringe displacements on each face are examined simultaneously. The first method will be discussed here and the second, which is now in use, in a subsequent publication.

In the experiments with the fringes formed by the two faces of the crystal, a GT cut quartz plate some 1·5 mm. thick was examined. The plate was worked plane on one surface and slightly concave on the other to give some eight rings across the whole surface. The large gap between the two silvered surfaces (1·5 mm.) leads to experimental difficulties. In effect the Fizeau system has the equivalent Fabry–Perot interferometer chromatic resolving power which is some 200,000. This is more than enough to resolve some hyperfine structure of the green mercury line and also excludes the use of the high intensity line-broadened sources now generally adopted for multiple-beam interferometry. Nor can fringes of equal chromatic order be used. It is necessary to employ a source of narrow line width of the type used for hyperfine structure studies, and the one adopted here is the inductance-stabilized type of vacuum mercury arc described by Tolansky (1946). The collimation of the light beam is also critical because of the big separation between the silvered surfaces.

The two silver films were used as electrodes and the quartz plate was clamped at the centre of the silvered faces by two phosphor bronze strips which hold the plate normal to the beam of parallel light and serve as connections to the oscillatory circuit, which in this instance was of the Pierce type. Such a clamping arrangement is frequent in electrotechnic uses of crystals.

Plate III (*a*) shows the fringes formed with the quartz plate at rest, the doubling of the fringes being due to the birefringence of the quartz, the two component fringes being plane polarized. Plate III (*b*) shows the fringe appearance when the plate is oscillating in its fundamental thickness shear mode at a frequency of 100 kc/s. The fringe broadening apparently is of a different nature to that shown by the surface deformations alone. The oscillating fringes do not exhibit the bright envelope of the fringes in the earlier pictures. Microphotometer measurements have been made but hyperfine structure complications make interpretation difficult as yet. The broadening is due to a change in *optical* path but further information is required before the effects due to a change in μ and a change in t can be separated. This can be done and is under consideration.

It may be noted that all these experiments with moving silvered surfaces offer a method for the study of thin metallic films when subject to dynamic forces. It is clear from the very existence of nodes and adjacent anti-nodes that the silver film is being subjected to stresses and is moving differentially through amplitudes which are up to ten times the film thickness. The persistent maintenance of narrow fringe width shows that the distortions do not affect the reflectivity of the film, but it is possible that the resistivity may change, and studies in this direction are being undertaken. This behaviour of the film clearly suggests that it is not continuous.

§7. SHATTERING OF SOME CRYSTALS

As the amplitudes of vibration of some of the quartz plates were increased, points were reached at which the plates shattered. The critical amplitude, which can be observed from the fringe motion, appears to depend on the cut of the crystal plate and the type of motion it is undergoing. Fortunately the system gives warning when fracture is imminent. In general when the vibration becomes dangerously excessive the fringes first blur due to onset of bodily motion of the quartz plate as a whole relative to the flat, and often a high-pitched audio-frequency note is heard. Furthermore the fringes frequently show that a marked bending of the whole crystal face is occurring. Beyond this there is danger of shatter taking place.

The shattering of quartz crystals due to excessive amplitude of oscillation is of course well known and earlier studies have been reported on the directions of the fractures and their possible relation to crystallographic directions. Plate III (*c*) shows a quartz plate which has broken after oscillating at excessive amplitudes (the fringes given by the separate pieces are visible). The important breaks are rectilinear, with directions inclined at specific angles to each other. This plate was an X cut bar with the Z axis running parallel to the long dimension. The directions of the fractures are inclined at approximately 36° to the Z axis.

Light has been thrown on the mechanism of shattering by the finding of an unusual effect upon one X cut crystal of the same dimensions as the former. This crystal had been polished on both sides with a local polish finish of the same grade as that of most of the crystals studied. It was set into oscillation and at some time (unobserved) during this oscillation an effect took place which resulted in a surprising change in the topography of both front and back faces of the crystal. One face is shown in Plate III (*d*) with fringes of the crystal when at rest. Illumination is with the green and two yellow mercury lines. The extraordinary feature is the appearance of a number of very clearly delineated parallel crystallographic slip lines. Areas with sharp discontinuous steps in level have appeared. Plate III (*e*) is an enlargement of a part of a slip region using higher fringe dispersion. Most step heights are some 500 A. There are other slip regions with slip lines at an angle of about 80° to these, but the particular dispersion here is not suitable to show these up very clearly.

The slip lines are effectively oriented in the same directions as the lines of fissure in the former crystal. It is of some considerable interest to point out that this X cut crystal with the slip bands can now withstand a much higher amplitude of oscillation than that at which the X cut in Plate III (*c*) shattered; in fact the power available has so far failed to shatter it. This would suggest that the slip released internal strains and produced a crystal of greater strength. This crystal, with its unusual surface discontinuity continues to oscillate as shown by Plate III (*f*) which is an oscillation picture. However, it is notable that the amplitude distribution is highly asymmetrical, being much less in the main slip region than elsewhere. This might well indicate local excessive damping in the slip region.

The opposite face of the same crystal is equally surprising, the 'rest' fringes being shown in Plate III (*g*). Several relatively deep hollows have appeared and they are not directly behind the slip bands on the other face. In addition there also exists a number of sharp discontinuous slip bands but on a smaller scale than on the opposite face. Steps on this face approximate to 150 A. and require a different fringe distribution to show them up. Since it has been established in this laboratory by interference methods that surface steps of even only 40 A. can still be seen in the ordinary microscope, we have re-examined the known distorted surfaces microscopically and indeed the steps can be readily identified by out-of-focus microscopy. An example is shown in Plate III (*h*). It would appear that the phenomenon has involved a rearrangement of sub-blocks of crystals. If one assumes that the crystal is composed of lineage blocks of the form suggested by Buerger, then an explanation is possible in terms of readjustment of lineage blocks following the true crystallographic glide or slip in other regions. It would seem that this crystal has been caught in an intermediate stage prior to break up and has in a sense frozen into it an indication of the break-up mechanism.

REFERENCES

DYE, W. D., 1932, *Proc. Roy. Soc.* A, **138**, 1.

OSTERBURG, H., 1929, *Proc. Nat. Acad. Sci., Wash.*, **15**, 892; 1932, *J. Opt. Soc. Amer.*, **22**, 19; 1933 a, *Phys. Rev.*, **43**, 819; 1933 b, *J. Opt. Soc. Amer.*, **23**, 30; 1934, *Rev. Sci. Instrum.*, **5**, 183.

TOLANSKY, S., 1946, *High Resolution Spectroscopy* (London: Methuen), p. 21; 1948, *Multiple-Beam Interferometry of Surfaces and Films* (Oxford: Clarendon Press), p. 1.

On the Spherical Aberration of Electron Emission Systems

By O. KLEMPERER and Y. KLINGER
Department of Physics, Imperial College, London

MS. received 28th July 1950

ABSTRACT. Spherical aberration coefficient and source diameter have been measured for a 'hairpin' emission system. Both have been found to be smaller than the corresponding values of various other emission systems.

In connection with some recent ray-tracing measurements in electron spectrometers, we have investigated the spherical aberration of a 'hairpin' emission system which served as a 'point' source in these experiments. Hairpin systems are now widely used in many electronic devices, for example in electron microscopes, electron diffraction cameras and x-ray point-focus tubes, hence the results of our measurements might be of some interest.

The measuring arrangement is shown in Figure 1. There, C represents a hairpin-shaped tungsten–molybdenum wire of 0·08 mm. thickness, G is a grid electrode with a circular aperture of $r_g = 0{\cdot}8$mm. radius and within a variable distance c from the tip of the hairpin. A is a tubular anode carrying a pepperpot-like diaphragm P arranged as shown in Figure 1 (*a*) in a section at right angles to the axis z. The apertures in P are 0·05 mm. in diameter, arranged in two perpendicular straight lines along the x and y axes respectively. Electron pencils emerging through the apertures in P are traced by means of the sliding fluorescent target T (Klemperer and Wright 1939, Klemperer and Mayo 1948).

The measured results apply to a range of cathode-to-grid spacings given by $0 < c < r_g$ which corresponds to a range of penetration percentages $3{\cdot}3\% > D > 0{\cdot}8\%$ (D is measured as the ratio of 'blackout' grid voltage to anode voltage). The grid electrode was connected to cathode potential or to a slightly negative bias. The filament heating current was kept sufficiently low (this was very necessary for small spacings) in order to avoid splitting up of the pencils (cf. Ellis 1947). This splitting up occurs spontaneously and is probably due to a generation of space charge oscillations (Klemperer 1947). Emission currents of the order of 100 μa. were drawn from the cathode during the measurements, anode voltages somewhere between 1 and 3 kv. being required according to the penetration percentage of the system investigated.

The following results were observed:

(i) The aberration coefficient $a = \Delta z/\Theta^2$ was found in all cases to be below 100 units where Δz, the distance between paraxial and zonal foci was measured in grid aperture radii r_g, and the semi-vertical angle of divergence Θ of the corresponding pencil was in radians.

(ii) The practically homocentric emission was completely symmetrical. No preferential directions of emission were observed (up to $\Theta = 0{\cdot}15$) either in the plane of the hairpin (y direction of Figure 1) or at right angles to it.

(iii) The source of emission appeared to be a circular disc of the order of one-tenth of a grid aperture radius or even less. Within the accuracy of the measurement (1 aperture radius) the emitting disc was located in the plane of the grid aperture.

Here, the size of the emitting source has been estimated from the divergence of the narrow pencil transmitted through the axial aperture of the ' pepperpot ' diaphragm. As shown in Figure 2, the emitting disc (probably the surface of a small space charge in front of the tip of the hairpin) may be located at J, the diaphragm at P and the target at T.

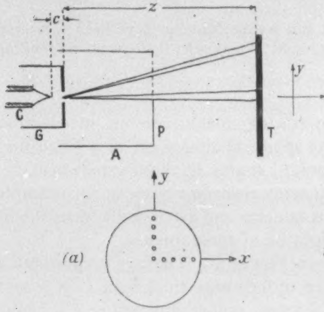


Figure 1. Tracing of pencils from hairpin emission system.

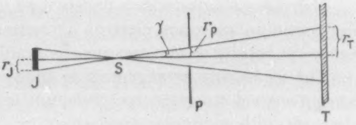


Figure 2. Selection of electrons by one aperture of the pepperpot diaphragm.

The radius of the ' pepperpot ' hole r_P and the radius of the fluorescent spot r_T can be measured with a microscope, and the semi-vertical angle of the pencil can be obtained as $\gamma = (r_T - r_P)/PT$, where PT is the distance between diaphragm and target. Again by simple geometry, the radius of the emitting disc is derived as $r_J = \gamma JP - r_P$ where JP is the distance between the diaphragm and the emitting disc J. The position of J, however, has been derived from tracing experiments as the intersection of rays which are given as the axes of the various pencils emerging from the 'pepperpot' holes.

Measurements of the size of the emitting disc cannot be expected to yield accurate results since, due to the thermal distribution of emission velocities, the intensity at the edge of the fluorescent spot decays exponentially and the boundaries are not sharply defined (cf. Klemperer 1947). However, the order of magnitude of r_J for a given system seems to be reproducible within a factor of two to three under different conditions of voltage and current.

It might be of some interest to compare the aberration and the radius of the emitting disc of the hairpin system with corresponding values obtained for some other emission systems.

In the following Table are given the relevant figures for five different emission systems—namely, the aberration coefficients *a* as defined above and the radii r_J of

System	T.D.L.	T.D.S.	O.A.	C.A.	H.P.
a/r_g	4000	<100	500	<100	<100
r_J/r_g	0·4	0·1	0·1	0·1	<0·1

the apparent electron source, all lengths being expressed in units of grid-aperture radii r_g. The first two systems—'T.D.L.' and 'T.D.S.'—are the two-diaphragm emission systems with large and with small spacing respectively between grid electrode and cathode as described by Klemperer and Mayo (1948). The next system, 'O.A.', represents an open-anode type, having a plane cathode, a plane grid electrode and a tubular anode without diaphragm. The fourth system on the list, 'C.A.', has a conical anode in front of a large concave cathode and has been described by Klemperer and Wright (1939). The last column contains information about the hairpin 'H.P.'

Comparison shows that the hairpin emission system is one of the best available electron sources. It should be recommended for electron optical apparatus if neither large currents nor an efficient modulation of the emitted intensity are required.

REFERENCES

ELLIS, S. G., 1947, *J. Appl. Phys.*, **18**, 879.
KLEMPERER, O., 1947, *Proc. Roy. Soc.* A, **190**, 376.
KLEMPERER, O., and MAYO, B. J., 1948, *J. Instn. Elect. Engrs.*, **95**, Pt. III, 135.
KLEMPERER, O., and WRIGHT, W. D., 1939, *Proc. Phys. Soc.*, **51**, 296.

Investigation of Electron Optical Properties of an Electrostatic Focusing System*

BY W. E. SPEAR
Birkbeck College Research Laboratory, University of London

MS. received 19th July 1950, and in final form 25th October 1950

ABSTRACT. It is shown that the combination of paraxial ray tracing and a dynamical determination of the electron trajectories in the cathode region leads to a complete analysis of electron-optical immersion systems, once the potential distribution is known. The method is applied to an electrode arrangement recently described for use in a fine focus x-ray tube. In particular it is shown that a demagnified image of the emitting area of the hairpin filament is produced close to the anode, if the outer electrode of the system is at cathode potential. A magnified image is formed near the emitter if this electrode is at anode potential. The results are in good agreement with experiment. A discussion of aberration and crossover is included.

§1. INTRODUCTION

A DETAILED investigation of electron-optical (e.o.) immersion systems presents certain difficulties not normally encountered with other elements, such as bipotential or unipotential lenses. There it is possible to isolate the focusing field and regard object and image planes as lying in field-free space. Moreover, the electrons entering the field are considered as homogeneous bundles of a definite velocity. Here, however, object and image plane are immersed in the e.o. field, and the trajectory of an electron emitted at an angle θ to the axis will depend strongly on the initial velocity v. It is the object of this paper to describe a method of obtaining the complete trajectories in this type of field which largely overcomes the above-mentioned difficulties and makes it possible to extend the investigation over a wide range of initial velocities and angles of emission.

The method is applied to study the properties of the focusing system shown in Figure 1. This consists of the grid cylinder G, the anode rod A, mounted coaxially with G, and the surrounding outer electrode E, at potential V_E, which forms an integral part of the system. The tip of the hairpin filament F lies on the axis of symmetry, in the plane of the aperture of the cathode shield C (aperture diameter 0·25 cm.). The H.T. is applied between A and G; the filament is kept at a small positive potential V_F with respect to the grid cylinder. The system was developed for use in a fine focus X-ray tube. Its experimental investigation and application have been described in a previous publication (Ehrenberg and Spear 1951). It was of interest in this connection to study the nature and formation of the electron concentration observed on the anode surface, and in particular the effect of the potential V_E on the focusing properties. Such an investigation involves (*a*) the determination of the potential distribution throughout the system and (*b*) the tracing of the electron trajectories through the field.

---

* This paper forms part of a thesis approved by the University of London for the Ph.D. degree

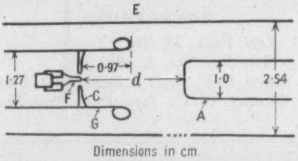


Figure 1. Section through focusing system.

§ 2. POTENTIAL DISTRIBUTION

The potential distribution was found by means of the relaxation method (Motz and Klanfer 1946, Luciow 1946) for the cases $V_E = V_G$ (System I), and $V_E = V_A$ (System II). The effect of space charge and the small geometrical and potential asymmetry introduced by the filament were neglected. At first a complete distribution was obtained for simplified geometry and boundary conditions, which for System I is shown in Figure 2. The grid aperture–anode distance corresponds to the case $d = 2·06$ cm., so that the anode surface lies close to the experimentally determined circle of least confusion. In the axial region the space was divided up by a relaxation mesh which had 25 nodes between G and A. In the following, one such division will be called one 'unit' (= 0·790 mm.). Unless otherwise stated all distances along the z axis will be measured in units. The

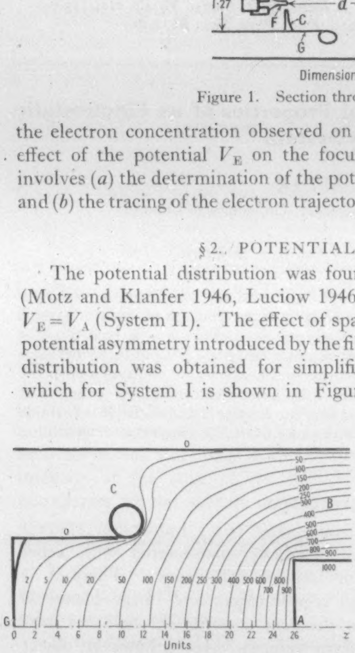


Figure 2. System I: Complete potential distribution. Anode potential 1,000. Boundary conditions: at C, $V=0$ in plane $z=8$; at B, potential decreases logarithmically towards outer electrode in plane $z=40$.

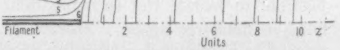


Figure 3. System I : Detailed potential distribution in grid cylinder. Anode potential 10,000.

anode potential was taken as 1,000 for the 1-unit mesh. The final accuracy of the potential at the nodes is then $\pm 0{\cdot}5$.

The 26-step mesh is too coarse to provide finer detail in the cathode region. Figure 3 illustrates the results of a more detailed investigation. A $\frac{1}{2}$-unit mesh has been introduced near to the grid aperture, which continues along the filament up to the plane $z = -4$. The potentials on plane $z = 12$, at the open end of the cathode cylinder, were taken as boundary values. All potentials given in Figure 3 refer now to an anode potential of 10,000. The filament is idealized to a thin circular cylinder, whose flat base lies in the plane $z = 0$. The radius of the cylinder is 0·1 unit, equal to that of the filament wire used in the experiments. An assumption of this kind is necessary, because the asymmetry of a real hairpin filament would have made the problem essentially a three-dimensional one, too complex for solution. The experimental results indicate that the asymmetry of the filament can only have a small effect on the properties of the system. The above assumption seems therefore justifiable. The filament has a positive bias of 6 with respect to the grid aperture; this corresponds to the experimentally determined value of 30 v., for a beam current of 300 μa., at 50 kv. The accuracy of the potential function in the cathode region is therefore $\pm 2{\cdot}5$ v. A further increase of the accuracy is hardly justifiable in view of the approximations made. Moreover, the ray tracing, ~~based~~ on the above potential distribution, gives results which show satisfactory agreement with experiment whenever a direct comparison is possible.

Figure 4 shows the complete potential distribution for System II. The geometry is identical with that of System I, but now both the outer tube and the anode are at potential 1,000. The region close to G is here at potential 40, as it was found experimentally that for optimum spot size $V_F = 200$ v. Graphs of the axial potentials in the two cases are given in Figure 5. The essential differences between the fields are: (*a*) the much stronger field penetration into the cathode cylinder of System II, mainly due to the large potential gradient between cathode cylinder and outer wall, and (*b*) the change in curvature of the equipotentials in System II about half-way between anode and filament.

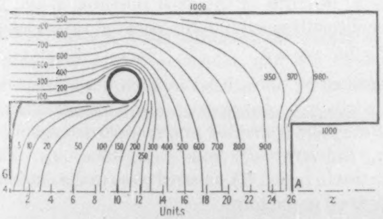


Figure 4. System II : Complete potential distribution. Anode potential 1,000.

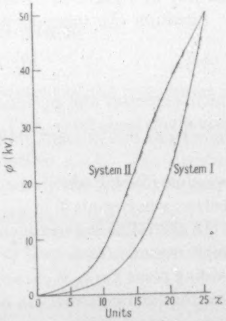


Figure 5. Axial potential in systems I and II.

The axial potential is of particular interest, and will be denoted by $\phi(z)$. The functions $a = \phi'/\phi$, and $b = \phi''/\phi$ can now be calculated at any node on the axis from the axial potential distribution and its differences. Provided that the interval Δz between successive nodes is sufficiently small, we obtain at $z = z_n$

$$a_n = \frac{\Delta\phi_{n-1} + \Delta\phi_n}{2\phi_n \Delta z}\,; \qquad b_n = \frac{\Delta^2 \phi_{n-1}}{\phi_n (\Delta z)^2}. \qquad \ldots\ldots(1)$$

These quantities have been tabulated for use in paraxial ray tracing. To normalize the potential as mentioned below in connection with equation (5) it is merely necessary to deduct the filament potential from the potential of every node and boundary in the relaxation pattern.

§ 3. DETERMINATION OF TRAJECTORIES

There are essentially two different approaches to the problem of determining the electron paths for a given potential distribution:

(i) By solving the time dependent equations of motion of the electron, which in the purely electrostatic case are

$$m\ddot{z} = -e\frac{\partial V}{\partial z}\,; \qquad m\ddot{r} = -e\frac{\partial V}{\partial r}\,, \qquad \ldots\ldots(2)$$

when $e = -4{\cdot}8 \times 10^{-10}$ E.S.U., and the z axis $(r = 0)$ forms the axis of symmetry. A general numerical method, based on this approach, has been developed by Goddard (1944), and applied by Goddard and Klemperer (1944) to the detailed investigation of a magnetic lens.

(ii) By a solution of the ray equation

$$r'' + \frac{1 + (r')^2}{2V}\frac{\partial V}{\partial z} r' - \frac{1 + (r')^2}{2V}\frac{\partial V}{\partial r} = 0, \qquad \ldots\ldots(3)$$

where r' and r'' denote differentiations with respect to z. In general, equation (3) can only be integrated by numerical step by step methods.

Equation (3), together with the well-known expansion

$$V(r, z) = \sum_{n=0}^{\infty} (-1)^n \frac{(\phi)^{(2n)}}{(n!)^2}\left(\frac{r}{2}\right)^{2n}, \qquad \ldots\ldots(4)$$

which expresses the potential $V(r, z)$ in terms of the axial potential ϕ and its derivatives, leads for $(r')^2 \ll 1$, and sufficiently small r, to the paraxial ray equation

$$4r'' + 2ar' + br = 0. \qquad \ldots\ldots(5)$$

Equation (5) will always be assumed to be normalized in such a way that the electron velocity $v = 0$, at $\phi = 0$.

In the following treatment, the e.o. field is divided into two parts: region (1) which extends from $z = 0$ to $z = z_m$ and region (2) from $z = z_m$ to $z = z_n$. The dividing plane $z = z_m$ is chosen such that in region (1), situated close to the emitting area, the axial potential can be expressed in the form

$$\phi = pz^2 + qz \qquad \ldots\ldots(6)$$

within the accuracy to which ϕ is known. This leads to an analytical solution of equations (2). Equation (5) is not generally applicable in this region on account of the large angles of emission and the differences in initial velocity. At $z = z_m$, however, most electrons have been accelerated sufficiently to make the effect of initial velocities quite negligible, and their inclination to the axis falls within paraxial limits. Equation (5) may therefore be applied in region (2) to determine the trajectories of electrons emitted over a wide range of initial conditions. In the following, relativistic effects are neglected, and only trajectories lying in a meridional plane are considered.

(i) *Analytic Solution in Region* (1)

Substitution of equation (6) into equation (4) leads to

$$V = pz^2 + qz - \tfrac{1}{2}pr^2. \qquad \ldots\ldots(7)$$

From equations (2) we obtain

$$\mathbf{m}\ddot{z} = -\mathbf{e}(2pz + q) \qquad \ldots\ldots(8)$$

and

$$\mathbf{m}\ddot{r} = \mathbf{e}pr. \qquad \ldots\ldots(9)$$

Let $l^2 = -(\mathbf{e}/\mathbf{m})p$, then

$$z = A\exp(\sqrt{2}lt) + B\exp(-\sqrt{2}lt) - q/2p \qquad \ldots\ldots(10)$$

and

$$\dot{z} = Al\sqrt{2}\exp(\sqrt{2}lt) - Bl\sqrt{2}\exp(-\sqrt{2}lt). \qquad \ldots\ldots(11)$$

The constants A and B are determined from the conditions $\dot{z} = \dot{z}_0$ at $t=0$, $z=0$ at $t=0$, giving

$$A = \frac{\dot{z}_0}{2l\sqrt{2}} + \frac{q}{4p}; \qquad B = \frac{\dot{z}_0}{2l\sqrt{2}} + \frac{q}{4p}. \qquad \ldots\ldots(12)$$

On the other hand, (9) leads to

$$r = (\dot{r}_0/l)\sin lt + r_0\cos lt \qquad \ldots\ldots(13)$$

and

$$\dot{r} = \dot{r}_0\cos lt - r_0 l\sin lt, \qquad \ldots\ldots(14)$$

where the conditions $\dot{r} = \dot{r}_0$ at $t=0$, and $r = r_0$ at $t=0$, serve to determine the values of the constants of integration. The calculation of r involves the value of lt, which has to be eliminated from (10):

$$lt = \frac{1}{\sqrt{2}}\ln\left\{\frac{z + q/2p + [z^2 + (q/p)z + \dot{z}_0/2l^2]^{1/2}}{\dot{z}_0/\sqrt{2}l + q/2p}\right\}. \qquad \ldots\ldots(15)$$

Hence, at any point z, for $0 < z \leqslant z_m$, r can be determined from equations (12) and (15).

r' is obtained from equations (11) and (14) by the relation

$$r' = \dot{r}/\dot{z}. \qquad \ldots\ldots(16)$$

Equations (13), (15) and (16) give a complete description of the trajectories in region (1).

For System I the tabulated differences of ϕ show that, within the known accuracy of ϕ, $\Delta^2\phi$ = constant, and $\Delta^3\phi = 0$ from $z=0$ to $z=3$. The dividing plane is therefore chosen at $z_m = 3$. In the following, only the values of r and r' in the plane $z=3$ will be required. Equations (11) and (15) show that $\dot{z}_3$ and lt_3 are functions of $\dot{z}_0$ only. Both vary slowly within the range concerned, and it is sufficiently accurate to obtain the values of z_3 and lt_3 by graphical interpolation between a few computed points.

(ii) *Solution in Region* (2)

The transformation of (r_3, r'_3) to a point in any plane $z = z_n$, by solutions of the paraxial ray equation, is based on two assumptions: (*a*) all electrons at $z=3$ have the same velocity, (*b*) the values of r_3 and r'_3 fall within the range of paraxial ray tracing. Assumption (*a*) is obviously justified because $\phi_3 = 285 \pm 2{\cdot}5$ v., and the most probable energy of emission is about 0·35 ev. The range of validity of (*b*) will be discussed in § 4.

Suppose now, $r = f_1(z)$ and $r = f_2(z)$ represent two independent solutions of (5) between $z=3$ and $z=z_n$ obtained by a step by step method of ray tracing, in which a and b are regarded as linear functions of zw ithin any interval (Maloff and Epstein 1938). Owing to the linearity of equation (5) any other solution can be written as a linear combination of the fundamental rays f_1 and f_2. Thus the radial height R_n and the gradient R'_n in any plane $z=z_n$ are given by

$$R_n=C(f_1)_n+D(f_2)_n, \qquad R'_n=C(f'_1)_n+D(f'_2)_n, \qquad \ldots\ldots(17)$$

where C and D are constants determined by the values of f_1 and f_2 at $z=z_m$. The values of C and D at $z_m=3$ are found to be

$$C=\frac{r'_3(f_2)_3-r_3(f'_2)_3}{(f_2)_3(f'_1)_3-(f'_2)_3(f_1)_3}; \quad D=\frac{r_3(f'_1)_3-r'_3(f_1)_3}{(f_2)_3(f'_1)_3-(f'_2)_3(f_1)_3}. \qquad \ldots\ldots(18)$$

For convenience, the initial conditions relating to the fundamental rays are chosen such that $(f_1)_3=0$, and $(f'_2)_3=0$. This leads to the following general transformations from plane $z=3$ to $z=z_n$:

$$R_n=(F_1)_n r'_3+(F_2)_n r_3, \qquad R'_n=(F'_1)_n r'_3+(F'_2)_n r_3, \qquad \ldots\ldots(19)$$

where $(F_1)_n=(f_1)_n/(f'_1)_3$; $(F_2)_n=(f_2)_n/(f_2)_3$ etc.

From the known values of the fundamental rays at $z=26$, the transformations on to the anode surface are found to be

$$R_{26}=1{\cdot}8012\,r'_3-0{\cdot}0380\,r_3, \qquad R'_{26}=-0{\cdot}1267\,r'_3-0{\cdot}0354\,r_3. \qquad \ldots\ldots(20)$$

(iii) *Transformation of Rays from off-axis Points*

Although the general transformations (19) apply to all rays in the meridional plane, irrespective of the value of r_0, it has been found convenient to discriminate between trajectories from the axial point, and those from off-axis points. The radial height R_n in region (2) of an electron emitted from a point r_0 mm. off the axis, with velocity v, and at an angle θ, can be expressed in terms of the radial height $(R_n)_0$ of an electron emitted with the same velocity and at the same angle from the axial point. Thus equations (13), (14) (16) and (19) lead to

$$\left.\begin{aligned} R_n&=(R_n)_0-r_0[(F_1)_n l\sin(lt_3)/\dot{z}_3-(F_2)_n\cos(lt_3)],\\ R'_n&=(R'_n)_0-r_0[(F'_1)_n l\sin(lt_3)/\dot{z}_3-(F'_2)_n\cos(lt_3)]. \end{aligned}\right\} \qquad \ldots\ldots(21)$$

§4. ACCURACY AND RANGE OF TRANSFORMATIONS

The accuracy of the fundamental rays depends (*a*) on the accuracy of the potential, and (*b*) on that of the step by step process by which the rays are traced. An estimate of (*b*) can be obtained by using a second independently determined set of fundamental rays in region (2), to derive f_1, f_2 and f'_1, f'_2 at a number of points along their trajectories. Comparison with the original values shows a maximum difference of 1–2% for the function and 3–4% for the gradients. As the rays are traced with approximately twice the accuracy warranted by the accuracy of the axial potential, (*a*) may be estimated as about 4% for the function and 8% for the gradients.

The range of validity of the transformations is limited by the paraxial conditions. If a further term of the expansion (4) is included in expressing V, $\partial V/\partial z$ and $\partial V/\partial r$ of equation (3) we obtain:

$$4r''+2a(1+\alpha)r'+b(1+\beta)r=0.$$

The correction terms α and β depend on $(r')^2$, r^2, a, b and higher derivatives of ϕ. Their effect on the step process has been investigated in the critical region from $z=6$ to $z=13$ and it is found that the error in r_{13} and r'_{13} is of the same order as that introduced by causes (*a*) and (*b*) for $\theta \simeq 45°$ and velocities up to $v = 10^8$ cm.sec^{-1}, (i.e. about three times the most probable velocity). These values therefore limit the range of the paraxial transformation.

Larger angles of emission can be treated with the same accuracy, provided that the velocity range is restricted. If this is not done, the determined curvature of the rays in region (2) will be too small. Hence the e.o. field tends to produce overfocusing (positive aberration) for large angles and velocities of emission.

§ 5. IMAGE FORMATION AND CROSSOVER IN SYSTEM I

(i) *Results of the Ray Tracing*

The trajectories given in Figures 6, 7 and 8 show some of the general results obtained. The velocities chosen correspond to a slow, a most probable, and a fast electron, emitted from points $r_0 = 0$ and $r_0 = 0{\cdot}0500$ mm. The angles of emission shown are $\theta = 6°$, 18°, 42°. The first gives rise to rays which fulfil the paraxial conditions all along their paths; those emitted at 18° and 42° will do so only in region (2). In addition, trajectories of electrons leaving the off-axis point at zero

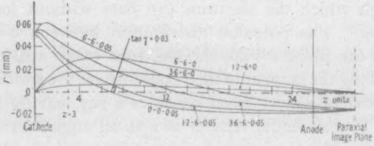


Figure 6. System I : Electron trajectories. $\theta = 6°$.

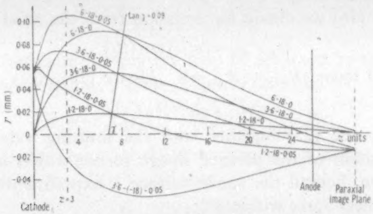


Figure 7. System I : Electron trajectories. $\theta = 18°$.

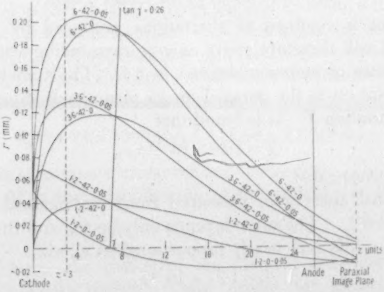


Figure 8. System I : Electron trajectories. $\theta = 42°$.

velocity and at a negative angle of emission are included. The rays are denoted by a group of three numbers; the first represents the velocity of emission ($10^{-7} v$ cm. sec^{-1}), the second the angle of emission, and the third gives the value of r_0.

The trajectories clearly show the formation of a demagnified image of the emitting area of the filament, situated behind the anode. It will be seen from Figure 8 that the e.o. field in region (1) underfocuses rays emitted at larger angles and velocities, giving rise to negative aberrations.

(ii) *The Paraxial Image Plane*

To derive transformations on to the paraxial image plane, it is necessary to extrapolate the fundamental trajectories beyond the plane $z=26$. For this purpose, the anode surface is assumed to consist of an infinitely thin conducting membrane, through which the electrons can pass without loss in velocity, or change in direction. The potential distribution between $z=0$ and $z=26$ is unchanged, but at the plane $z=26$, the electrons enter a field-free region and continue to move along the tangent to their trajectory at the anode plane. In order to locate the paraxial plane, we must consider a ray leaving the centre of the emitting area with a small initial velocity at a small angle. As $\dot{z}_0 \to 0$ in (15), lt_3 and therefore $\dot{z}_3$ (see (11)), approach limiting values, which change very slowly with $\dot{z}_0$ for small values of $\dot{z}_0$.

From (13) and (16) we obtain for emission from the axial point

$$(r_3/r'_3)_0 = (\dot{z}_3/l) \tan (lt_3). \qquad \ldots\ldots(22)$$

But in the paraxial focus $(R_n)_0 = (R_p)_0 = 0$. Hence (19) gives

$$-(F_1/F_2)_p = (r_3/r'_3)_0. \qquad \ldots\ldots(23)$$

Equations (22) and (23) together with the limiting values of lt_3 and $\dot{z}_3$ determine the position of the paraxial image plane, which is found to lie at $z=29{\cdot}568$, 2·82 mm. behind the anode surface. Experimental and theoretical object–image distances agree within 4%.

(iii) *Aberrations in the Image Plane*

This treatment is confined to aberrations produced by the e.o. field in region (1), and it will therefore apply only to rays falling within the range of angles and velocities of emission defined in § 4. The transformation of rays from the axial point on to the paraxial image plane leads directly to a value of the transverse aberration T. It is found that

$$T = 2{\cdot}34 \times 10^{-18} \dot{z}_0 \dot{r}_0 \qquad \ldots\ldots(24)$$

within the range investigated.

The longitudinal aberration, measured from the paraxial image plane, is given by $L = -(R/R')_p$. The ratio depends only on lt_3, and therefore $\dot{z}_0$, and is independent of $\dot{r}_0$. L is very nearly proportional to $\dot{z}_0$, viz.:

$$L = 1{\cdot}37 \times 10^{-8} \dot{z}_0. \qquad \ldots\ldots(25)$$

(iv) *The Crossover*

The existence of a crossover in immersion lenses was first pointed out and explained by Maloff and Epstein (1934). More recently Einstein and Jacob (1948), described an investigation of the properties of the crossover in an immersion lens of the conventional type, using a flat cathode. The treatment, however, is restricted to paraxial rays and comparatively small velocities of emission; the plane of the crossover is assumed to be perpendicular to the axis, passing through the point of intersection of rays which leave the emitting area at $\phi=0^\circ$ and with a given velocity v.

The methods developed here lend themselves to an investigation over a far wider range of angles and velocities. It has been found more convenient to generalize the above definition and regard the crossover as the locus of the intersection of rays, leaving different points of the emitting surface at the same angle and with the same velocity. Thus the condition $R_n=(R_n)_0$ determines a point on the crossover for any given v and θ. From (21) and (22) the ratio $(F_1/F_2)_c$ at the crossover is found to be:

$$(F_1/F_2)_c=(\dot{z}_3{}^2/l^2)\,(r'_3/r_3)_0. \qquad \ldots\ldots.(26)$$

The crossover abscissa is thus independent of r_0 and $\dot{r}_0$ but depends (via $\dot{z}_3$) on lt_3, and therefore on $\dot{z}_0$. Hence, a longitudinal aberration L_c is to be expected, which is clearly shown in Figures 6, 7 and 8.

It was found that F_1/F_2 could be represented with sufficient accuracy in the region of the crossover by a linear equation. For $\dot{z}_0=0$, $z_c=6{\cdot}933$ units; this point is now taken as the origin from which L_c is measured in the positive z direction. It is proportional to $\dot{z}_0$, and given by

$$L_c=1{\cdot}64\times10^{-8}\,\dot{z}_0. \qquad \ldots\ldots(27)$$

The crossover ordinate for any ray may be obtained by transformation into its particular crossover plane. Thus (19) and (26) lead to

$$R_c=(F_2)_c(r_3)_0[1+(\dot{z}_3{}^2/l^2)(r'_3/r_3)_0{}^2]. \qquad \ldots\ldots(28)$$

F_2 varies slowly in the region of the crossover, and a mean value was used to calculate the values of R_c. It is found that

$$R_c=4{\cdot}83\times10^{-9}\,\dot{r}_0. \qquad \ldots\ldots(29)$$

It follows from (27) to (29) that for a given angle of emission θ the crossover of all trajectories will lie on a conical surface with apex at $(z_c, 0)$, whose semi-vertical angle γ (see Figures 6, 7, 8) depends only on θ, and is given by $\tan\gamma=0{\cdot}294\tan\theta$. On the other hand, the locus for constant velocities of emission will be the surface of an ellipsoid of revolution, whose section is

$$R_c{}^2/(4{\cdot}83\times10^{-9}v)^2+L_c{}^2/(1{\cdot}64\times10^{-8}v)^2=1.$$

It is of interest to compare crossover and image size. The latter is given by $R_p=Mr_0$, where M is the magnification of the system. It follows then, that

$$R_c/R_p=4{\cdot}83\times10^{-9}\,\dot{r}_0/Mr_0. \qquad \ldots\ldots(30)$$

Taking $r_0=0{\cdot}075$ mm., the radius of the filament wire used, and $M=0{\cdot}32$ (see §7), $|R_c/R_p|=1{\cdot}9\times10^{-7}\,\dot{r}_0$. Thus, except for electrons leaving with extremely small initial velocities (of the order of $0{\cdot}5\times10^7$ cm. sec^{-1}), which form only a negligible portion of the total emission, the image in this system will always be smaller than the crossover. The usual definition of the crossover as the 'minimum section' of the beam (e.g. von Borries and Ruska 1939, Cosslett 1946), does not therefore apply in this case. Einstein and Jacob (1948) have investigated an immersion lens of the type used in cathode-ray tubes ($M\simeq2$, $r_0\simeq5$ mm.).

Their graphs lead to $R_c = 18 \times 10^{-9} \dot{r}_0$. Using these values as in (30) $|R_c/R_p| = 0{\cdot}018 \times 10^{-7} \dot{r}_0$ which shows that for practically all velocities this type of immersion lens produces a crossover considerably smaller than the image. The crossover now forms the minimum section of the beam and, as Maloff and Epstein point out, a smaller spot will be observed on the final screen of the tube if the second lens images the virtual crossover, and not the cathode image.

§ 6. RESULTS FOR SYSTEM II

The trajectories, given in Figure 9, illustrate the image formation in System II. They were obtained by the method described in § 3. The paraxial image plane lies now inside the grid cylinder where an image with magnification 1·6 is formed. The crossover, situated at about $z = 0{\cdot}3$, represents here the minimum section of the beam. On the whole, the image formation in System II resembles closely that of the more conventional type of immersion lens mentioned above.

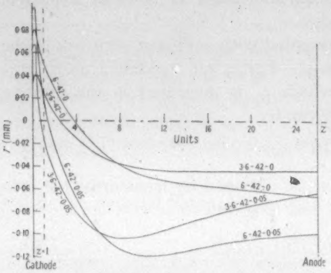


Figure 9. System II : Electron trajectories. $\theta = 42°$.

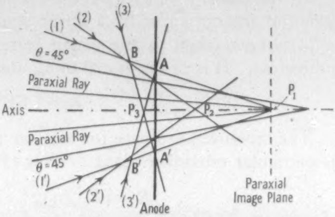


Figure 10. Diagram illustrating formation of electron concentration on anode surface in presence of high angle electrons.

The aberrations in the image plane are much stronger than those of System I, particularly the longitudinal error. The rays, having passed the image plane, are now bent towards the axis, forming a disc on the anode surface, whose size depends largely on the velocity and angle of emission—much more so than the true image formed in System I. The results fully explain the experimental observations, and confirm that e.o. systems of this kind are unsuitable for use in a fine focus X-ray tube.

§ 7. COMPARISON OF EXPERIMENTAL AND THEORETICAL RESULTS

The general transformation (21) leads directly to a value for the demagnification in the paraxial image plane. As a ray (v, θ, r_0) is imaged at a distance $|R_p| + (R_p)_0$ below the image of the axial ray $(v, \theta, 0)$, M is given by

$$|M| = [\,|R_p| + (R_p)_0]/r_0 = (F_1)_p l \sin(lt_3)/\dot{z}_3 + (F_2)_p \cos(lt_3). \qquad \ldots\ldots(31)$$

In System I, M changes very little with angle and velocity of emission and the trajectories investigated give a mean value of $M = 0{\cdot}32$ in the paraxial image plane. The same applies to the size of the spot on the anode surface. Here the mean value of M is found to be 0·33, about 3% higher than in the paraxial image plane.

On the assumption that the diameter of the emitting area is equal to the diameter of the filament wire used (0·006 in.), $2r_0 = 150\mu$; this leads to a spot size on the anode surface of $2R_{26}=50\mu$, in good agreement with the experimental value of the half-width for $d=2{\cdot}06$ cm.

It is of interest in this connection to obtain values of the magnification, using the well-known theorem of Helmholtz–Lagrange, which may be written in the form

$$M=v\sin\theta/v_i\sin\theta_i, \qquad \ldots\ldots(32)$$

where v_i and θ_i denote velocity and angle of convergence at the image plane. Sin θ_i can here be related directly to v and θ. If the limiting values of $\dot{z}_3$ and lt_3 (§ 5 (ii)) are then substituted, (32) will predict the magnification for electrons satisfying the paraxial conditions, which is found to be in good agreement with the results of equation (31).

For other rays the comparison shows that, within the range of trajectories treated, aberrations do not affect the image formation to any large extent. Clearly, it is not possible to account here for the small intensity thrown into the tails of the experimental distribution curves. These appear to be primarily due to electrons emitted at angles outside the range of this investigation, for which the transformation in regions (2) is no longer valid, or is only approximately true. Such electrons will be overfocused by the field in region (2) and it follows that $T=0$, not at $\theta=90°$ as predicted by (24), but at $\theta=\theta'$, where $\theta'<90°$. If $\theta>\theta'$, T will become negative and the system will produce a positive transverse aberration. Figure 10 illustrates the path of a number of axial rays near the anode surface. For rays (1) emitted at θ', $T=0$, and therefore an image point P_1 is formed on the paraxial plane. Rays (2), emitted at $\theta>\theta'$, cut the axis at P_2 between paraxial image plane and anode surface, whilst those electrons leaving at still larger angles, rays (3) and (3′), form an image in front of the anode surface. These rays will define a disc of least confusion AA′. Moving the anode surface from its optimum position AA′ towards the paraxial focus increases the intensity in the 'tails', but at the same time reduces the size of the intense central portion. The effect of moving the anode the opposite way should be slight, up to the disc BB′. Both these deductions are borne out by the experimental results.

ACKNOWLEDGMENT

The author wishes to express his gratitude to Professor J. D. Bernal for permission to work at the Birkbeck College Research Laboratory and to Dr. W. Ehrenberg for his continued advice and encouragement.

REFERENCES

von Borries, B., and Ruska, E., 1939, *Z. tech. Phys.*, **20**, 225.

Cosslett, V. E., 1946, *Introduction to Electron Optics* (Oxford : University Press), §47.

Ehrenberg, W., and Spear, W. E., 1951, *Proc. Phys. Soc.* B, **64**, 67.

Einstein, P. A., and Jacob, L., 1948, *Phil. Mag.*, **39**, 20.

Goddard, L. S., 1944, *Proc. Phys. Soc.*, **56**, 372.

Goddard, L. S., and Klemperer, O., 1944, *Proc. Phys. Soc.*, **56**, 378.

Luciow, B. A., 1946, *Ph. D. Thesis*, University of London.

Maloff, I. G., and Epstein, D. W., 1934, *Proc. Inst. Radio Engrs.*, N.Y., **22**, 1386; 1938, *Electron Optics in Television* (New York : McGraw-Hill).

Motz, H., and Klanfer, L., 1946, *Proc. Phys. Soc.*, **58**, 30.

Electron-Optical Systems with Helical Axis

By D. GABOR
Department of Electrical Engineering, Imperial College, London

MS. received 5th September 1950

ABSTRACT. As the chromatic error of focus has the same sign in all electron lenses with straight axis, achromatic combinations do not exist. It is shown, however, that systems with curved axes can be designed which are equivalent to centred systems in the gaussian approximation, but have chromatic error of opposite sign, and thus may be useful as correcting elements. Systems with a helical axis, with a helical electrostatic field so shaped as to eliminate first-order astigmatism, have particular advantages. In such systems images of unit magnification are produced at equal intervals along the helical axis. Under certain conditions the transverse chromatic shift can be made zero in every second, or third, . . . image, while the longitudinal error of focus can be given any negative value by suitable choice of the pitch. Such systems may find application in secondary emission and dark-field electron microscopes, and particularly in diffraction microscopy.

At a certain pitch the chromatic error becomes infinite. This system is particularly easily realized by means of two coaxial cylindrical electrodes. It is useless as a chromatic corrector, but may find applications as a velocity spectrograph.

§1. INTRODUCTION

It is well known (Scherzer 1936) that all electron lenses with a straight axis have chromatic errors of the same sign: the faster electron is always focused *beyond* the slower one. It is also known that electron mirrors have errors of opposite sign, but their application for chromatic correction meets with serious objections (Ramberg 1949). Apart from difficulties of design, centred electron mirrors have the great practical disadvantage that at the point of return the electrons are slowed down almost to a standstill, and are therefore very susceptible to unintentional disturbances of the electric field, such as may arise from polarization layers at the electrodes.

In ordinary transmission electron microscopy the unavoidable chromatic error of the objectives has no serious consequences. At the small apertures used at present the depth of focus is so large that sufficiently sharp focusing can be assured by careful stabilization of the voltage and current supplies. Even at the highest resolutions attained up to now it is sufficient to keep the fluctations within a few volts; the unavoidable energy spread of the beam electrons due to their thermal origin is about an order of magnitude lower and hence negligible. It is true that this eliminates only a part of the chromatic error; there remains the part which is due to energy losses of the beam electrons in the object, of the order of 15–25 volts per collision. But it has been shown (Gabor 1944, p. 32) that this part of the chromatic error is *beneficial* for the contrast in the bright-field, transmission method, and does not affect the resolving power. The advantage is turned into its opposite in dark-field illumination. A dark-field microscope operates chiefly with scattered electrons which have suffered losses, and requires an achromatic objective. This explains the inferior performance of the dark-field method in electron microscopy. The same is true of microscopes operating with secondary electrons, or with a photo-emitting object.

The need for achromatic lenses has become acute recently by the introduction of the method of ‘ diffraction microscopy ’ (Gabor 1949). In this method the aberrations of electron objectives are compensated by light-optical means in a reconstruction process carried out with visible light. This, however, is true only of those aberrations which distort, but do not destroy, the information contained in the diffraction diagram which forms the starting point of the reconstruction process. But chromatic aberration destroys some of the information by wiping out the more closely spaced interference fringes. Appreciable progress beyond the present stage of resolving power may be achieved by orthodox means: careful stabilization of voltages and currents, or the use of unipotential 'fixed focus' electrostatic lenses. But ultimately even the thermal energy spread of the order of 0·3 ev. will become a hindrance to increased resolution, unless the electron-optical system is, at least approximately, achromatized. Moreover, achromatic lenses might make it possible to use autoelectronic emission ('point emission'), which allows the production of current densities 10^3–10^4 times higher than in thermal emission, and thus shortens exposure times very appreciably. At present this is not feasible because of the energy spread, which is of the order of 5 volts in autoelectronic emission.

There is therefore sufficient practical inducement to take up the almost abandoned problem of achromatic lenses. A suggestion for the start is given by electron mirrors. In these the straight axis suffers a sudden break of 180°. This suggests that the required effect might be achieved by a gradual, instead of a sudden, break, and points to the study of systems with curved optic axis. A difficulty arises at the outset from the fact that such systems will not in general possess the property which in ordinary lenses is a direct consequence of the rotational symmetry, viz. they will not be free from astigmatism on the axis, and electrons of different velocity will not in general be refocused on the axis. But it can be shown by simple considerations that these difficulties can in fact be overcome.

§2. A SYSTEM WITH CIRCULAR AXIS

Electron-optical systems with curved axis have been studied by Cotte (1938) Wendt (1943) and Hutter (1948). Wendt in particular has described a system with circular axis which has zero astigmatism for electrons of a certain energy starting from a point of the axis which can be conveniently taken as a basis for an elementary explanation. It will be shown—a point not noticed by previous authors—that this system has, in fact, the properties required of a chromatic corrector.

It is well known that an electron beam which is forced into a circular path between two cylindrical electrodes is focused at intervals of $\pi/\sqrt{2}=127°$ in the plane of rotation, while it spreads out freely in the direction of the axis. If, now, following Wendt, one curves the electrodes into hyperbolic shape, as shown in Figure 1 (*a*), one can achieve equal focusing in both planes, i.e. eliminate the astigmatism on the circular axis c, but the beam will now be focused at intervals of 180°, as shown in Figure 1 (*b*). But this is true only of electrons of the same energy as the one which circulates on the axis. An electron with excess initial kinetic energy, which may be a fraction ϵ above that of the 'standard' electrons, will also describe an orbit which is circular in first approximation, but it will be eccentric, and of larger radius. We will show now that the excess energy electron will describe this orbit e with a velocity *below* the standard, because the mean potential on e is by *twice* the fraction ϵ below that on c. The mean potential on e is the same as the potential on the circle m, which is drawn with the same radius as e, but concentric with c. Let ϕ be the potential and V the volt-energy of the electrons. For a standard electron let $V=\phi_c$, the potential on the circular axis, with radius r. For the excess energy electron $V=(1+\epsilon)\phi_c$, and its curvature radius r_e can be calculated at the point P where it is tangential to c from the equation

$$\frac{1}{r_e} = -\frac{1}{2V}\frac{\partial\phi}{\partial r}. \qquad \ldots\ldots(1)$$

Substituting $V=(1+\epsilon)\phi_c$, and taking into consideration that r is an equilibrium orbit for $\epsilon=0$, we obtain, with the notation $r_e-r=\Delta r$,

$$\frac{\Delta r}{r}=\epsilon \quad \text{and} \quad \frac{\partial\phi}{\partial r}\Delta r=-2\epsilon\phi_c. \qquad \ldots\ldots(2)$$

This proves the statement that the mean potential on e is below the standard by a fraction 2ϵ, and as the kinetic energy at the point P was above the standard by a fraction ϵ, the mean velocity of the e-electron is in the mean below the standard by a fraction $\frac{1}{2}\epsilon$.

Electrons e′, e″ with the same energy as e will be focused around the e-axis in a similar way to the focusing of the standard electrons around c, but they do not arrive at the same point. It is easy to see, and will be proved formally later, that they perform oscillations around the e-axis with the same period as the standard electrons, at least to an approximation in which ϵ^2 can be neglected. Thus they come to a focus, but behind P for two reasons: first because, as we have seen, the faster electron in fact runs more slowly, and second because its orbit is longer, in the ratio ϵ, as may be seen from the first of equations (2). Thus they will rejoin the axis c in a point P‴ which is behind P by a distance

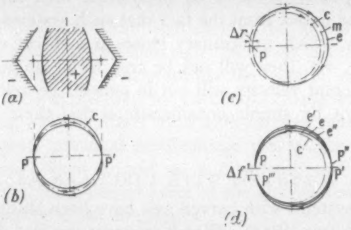


Figure 1. Elementary explanation.
(*a*) Electrode system, Wendt 1943. (*b*) Focusing of electrons with standard energy: c is the circular axis. (*c*) The axis e for electrons with excess energy. (*d*) Focusing of electrons with excess energy.

$$f=\tfrac{3}{2}\epsilon \times \text{length of the circular axis.} \qquad \ldots\ldots(3)$$

It is evident from Figure 1 (*d*) that it is only the *second* image of P, P‴ which we can use, not the first P″, where the focal points corresponding to different energies are strung out radially. But while the foci of even order only may be used for chromatic correction, the odd foci may be useful for velocity spectrography, as will be shown in more detail later on.

The arrangement with circular axis can be used in fact for the second purpose only, for, as the second focus nearly coincides with the initial point, the initial and the final beams could not be separated. This suggests twisting the axis into a helix. It will now be shown that the negative chromatic error not only retains its sign, but even increases beyond any limit within a certain range of pitch.

§ 3. THE EQUATIONS OF MOTION IN A HELICAL ELECTROSTATIC FIELD

Any combination of coaxial helical electrodes with the same pitch produces a helical field, and any helix in this field can be the trajectory of an electron with an appropriate velocity. It may be noted that a helical field is produced also if one of the electrodes is a circular cylinder, and that helical motion is possible also between two coaxial cylinders.

In cylindrical coordinates z, r, θ the helical axis may have the equations

$$r=r_0 \quad z=p\theta \qquad \ldots\ldots(4)$$

where p is the ' diametral pitch ', equal to the linear pitch divided by 2π. It will be convenient to adopt ' helical ' coordinates u, w such that

$$u=r-r_0, \quad w=\frac{r_0}{(p^2+r_0^2)^{1/2}}(z-p\theta)=\cos\alpha\,(z-p\theta), \qquad \ldots\ldots(5)$$

retaining θ as the third coordinate. As shown in Figure 2, du is in the direction of the radius of curvature, whose value is $r_0/\cos\alpha$, and dw is in the direction of the binormal, both at right angles to the helical axis. Any helix of pitch p has equations $u=$ const., $w=$ const. If u and w are small they measure the distance of a point in two directions at right angles from the central helix.

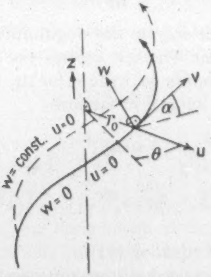


Figure 2. Helical coordinates.

The Lagrangian function of a (non-relativistic) electron is, in these coordinates,

$$\begin{aligned} L &= \tfrac{1}{2}\boldsymbol{m}v^2-\boldsymbol{e}\phi=\tfrac{1}{2}\boldsymbol{m}(\dot{z}^2+\dot{r}^2+r^2\dot{\theta}^2)-\boldsymbol{e}\phi \\ &= \tfrac{1}{2}\boldsymbol{m}\left[\left(p\dot{\theta}+\frac{(p^2+r_0^2)^{1/2}}{r_0}\dot{w}\right)^2+\dot{u}^2+(r_0+u)^2\dot{\theta}^2\right]-\boldsymbol{e}\phi, \qquad \ldots\ldots(6) \end{aligned}$$

where $\boldsymbol{e}$ is the electron charge and $\boldsymbol{m}$ the mass. The Lagrangian equations of motion are

$$\boldsymbol{m}\ddot{u}=-\boldsymbol{e}\frac{\partial\phi}{\partial u}+\boldsymbol{m}(r_0+u)\dot{\theta}^2, \qquad \ldots\ldots(7)$$

$$\boldsymbol{m}\frac{p^2+r_0^2}{r_0^2}\ddot{w}=-\boldsymbol{e}\frac{\partial\phi}{\partial w}-\boldsymbol{m}\frac{p}{r_0}(p^2+r_0^2)^{1/2}\ddot{\theta}, \qquad \ldots\ldots(8)$$

$$[p^2+(r_0+u)^2]\dot{\theta}+\frac{p}{r_0}(p^2+r_0^2)^{1/2}\dot{w}=C, \qquad \ldots\ldots(9)$$

where C is a constant. The last equation is the momentum integral relative to θ. The energy integral, for an electron with an energy excess a fraction ϵ above the standard electrons, is

$$\tfrac{1}{2}\boldsymbol{m}\left[\left(p\dot{\theta}+\frac{(p^2+r_0^2)^{1/2}}{r_0}\dot{w}\right)^2+\dot{u}^2+(r_0+u)^2\dot{\theta}^2\right]+\boldsymbol{e}\phi=-\boldsymbol{e}\epsilon\phi_0, \quad \ldots\ldots(10)$$

where ϕ_0 is the potential on the helical axis. For a standard electron which moves on the axis this gives the angular velocity $\dot{\theta}=\Omega$

$$\Omega^2=-\frac{2\boldsymbol{e}}{\boldsymbol{m}}\frac{\phi_0}{p^2+r_0^2}. \quad \ldots\ldots(11)$$

We need the momentum integral (9) for eliminating $\dot{\theta}$ and $\ddot{\theta}$ from the equations of motion. For this we must first determine the constant C. For the present investigation it will be sufficient to consider only electrons which at some point have crossed the axis. At this point $u=w=0$, $\phi=\phi_0$, so that the energy integral becomes

$$(p^2+r_0^2)\dot{\theta}^2+2\frac{p}{r_0}(p^2+r_0^2)^{1/2}\dot{\theta}\dot{w}+\dot{u}_0^2+\frac{p^2+r_0^2}{r_0^2}\dot{w}^2=-\frac{2\boldsymbol{e}}{\boldsymbol{m}}\phi_0(1+\epsilon), \quad \ldots\ldots(12)$$

while the momentum integral (9), squared and divided by $p^2+r_0^2$, is now

$$(p^2+r_0^2)\dot{\theta}^2+2\frac{p}{r_0}(p^2+r_0^2)^{1/2}\dot{\theta}\dot{w}+\frac{p^2}{r_0^2}\dot{w}_0^2=C^2/(p^2+r_0^2). \quad \ldots\ldots(13)$$

Subtraction gives $$C^2/(p^2+r_0^2)=-\frac{2\boldsymbol{e}}{\boldsymbol{m}}\phi_0(1+\epsilon)-(\dot{u}_0^2+\dot{w}_0^2), \quad \ldots\ldots(14)$$

which shows that C depends only on the longitudinal component of the velocity of the electron at the point where it crosses the axis, not on its transverse components $\dot{u}_0$, $\dot{w}_0$. To simplify we write ϵ' for the longitudinal excess fraction, and obtain the momentum integral in the form

$$\left\{[p^2+(r_0+u)^2]\dot{\theta}+\frac{p}{r_0}(p^2+r_0^2)^{1/2}\dot{w}\right\}^2$$
$$=-\frac{2\boldsymbol{e}}{\boldsymbol{m}}\phi_0(1+\epsilon')(p^2+r_0^2)=(1+\epsilon')(p^2+r_0^2)^2\Omega^2, \quad \ldots\ldots(15)$$

where we have made use of equation (11).

So far the equations are exact, but from this point on we neglect all powers of $u, w, \dot{u}, \dot{w}$ and ϵ higher than the first, and treat also $\theta-\Omega$ as a small quantity. In this *gaussian approximation* the last equation simplifies to

$$2r_0\Omega u+(p^2+r_0^2)(\dot{\theta}-\Omega)+\frac{p}{r_0}(p^2+r_0^2)^{1/2}\dot{w}=\tfrac{1}{2}\epsilon\Omega(p^2+r_0^2), \quad \ldots\ldots(15')$$

which can be used to eliminate $\dot{\theta}$ and $\ddot{\theta}$ in equations (7) and (8). In these equations we introduce, to the same approximation,

$$\frac{\partial\phi}{\partial u}=\phi_u+\phi_{uu}u+\phi_{uw}w, \qquad \frac{\partial\phi}{\partial w}=\phi_w+\phi_{wu}u+\phi_{ww}w, \quad \ldots\ldots(16)$$

where we have used the suffix notations for the derivatives at the helical axis, to distinguish them from those at an arbitrary point. Substituting (15′) and (16) into (7) and (8) we obtain zero-order terms and first-order terms at the right-hand sides. If we postulate that the helical axis is itself a trajectory (for standard electrons), we obtain the two conditions for the vanishing of the zero-order terms:

$$\phi_w = 0, \qquad \ldots\ldots(17)$$

$$\boldsymbol{e}\phi_u = \boldsymbol{m} r_0 \Omega^2. \qquad \ldots\ldots(18)$$

The first expresses that the gradient in the direction of the binormal must be zero. The second, combined with equation (11), can be written in the form

$$\left(\frac{1}{\phi}\frac{\partial\phi}{\partial r_0}\right) = -\frac{2r_0}{p^2+r_0^2} = -\frac{2}{R}, \qquad \ldots\ldots(19)$$

where R is the radius of curvature of the helix, by a well-known formula. Thus condition (18) expresses the balance of the electric and the centrifugal force. The first-order equations of motion now become

$$\ddot{u} = -\frac{\boldsymbol{e}}{\boldsymbol{m}}(\phi_{uu}u + \phi_{uw}w) + \frac{p^2-3r_0^2}{p^2+r_0^2}\Omega^2 u - \frac{2p\Omega}{(p^2+r_0^2)^{1/2}}\dot{w} + \epsilon r_0 \Omega^2, \qquad \ldots\ldots(20)$$

$$\ddot{w} = -\frac{\boldsymbol{e}}{\boldsymbol{m}}(\phi_{wu}u + \phi_{ww}w) + \frac{2p}{(p^2+r_0^2)^{1/2}}\dot{u}. \qquad \ldots\ldots(21)$$

These equations are of the form

$$\left.\begin{aligned} \ddot{u} &= -2\omega\dot{w} - \alpha_1 u - \beta w + \epsilon r_0 \Omega^2, \\ \ddot{w} &= +2\omega\dot{u} - \alpha_2 w - \beta u, \end{aligned}\right\} \qquad \ldots\ldots(22)$$

where we have put
$$\omega = p/(p^2+r_0^2)^{1/2}. \qquad \ldots\ldots(23)$$

Consider first standard energy electrons, $\epsilon = 0$. The equations (22) will reduce to the well-known equations of rotationally symmetrical electromagnetic electron lenses if, and only if,

$$\beta = \phi_{uw} = 0 \qquad \ldots\ldots(24)$$

and if $\alpha_1 = \alpha_2 = \alpha$, i.e.

$$\alpha = \frac{\boldsymbol{e}}{\boldsymbol{m}}\phi_{ww} = \frac{\boldsymbol{e}}{\boldsymbol{m}}\phi_{uu} - \frac{p^2-3r_0^2}{p^2+r_0^2}\Omega^2 = \frac{\boldsymbol{e}}{\boldsymbol{m}}\left(\phi_{uu} + 2\frac{p^2-3r_0^2}{p^2+r_0^2}\phi_0\right). \qquad \ldots\ldots(25)$$

The first condition postulates that the gradient in the direction of the binormal vanishes not only at the helical axis but also in its immediate neighbourhood on the helical surface $w = 0$. This is automatically satisfied if this surface is a surface of symmetry. The second is the condition of astigmatism. It postulates one relation between the second-order derivatives; a second relation is obtained from Laplace's equation.

Let us write for the potential

$$\psi(z, r, \theta) = \phi(u, w). \qquad \ldots\ldots(26)$$

Laplace's equation is, in cylindrical coordinates,

$$\frac{\partial^2\psi}{\partial z^2} + \frac{\partial^2\psi}{\partial r^2} + \frac{1}{r}\frac{\partial\psi}{\partial r} + \frac{1}{r^2}\frac{\partial^2\psi}{\partial\theta^2} = 0. \qquad \ldots\ldots(27)$$

Equation (26) with the definitions (5) gives the relations

$$\frac{\partial^2\psi}{\partial z^2} = \frac{r_0^2}{p^2+r_0^2}\frac{\partial^2\phi}{\partial w^2}, \quad \frac{\partial\psi}{\partial r} = \frac{\partial\phi}{\partial u}, \quad \frac{\partial^2\psi}{dr^2} = \frac{\partial^2\phi}{\partial u^2}, \quad \frac{\partial^2\psi}{\partial\theta^2} = \frac{p^2 r_0^2}{p^2+r_0^2}\frac{\partial^2\phi}{\partial w^2}, \qquad \ldots\ldots(28)$$

if ϕ does not depend explicitly on θ. Thus Laplace's equation for the helical field is

$$\frac{\partial^2\phi}{\partial u^2}+\frac{1}{r_0+u}\frac{\partial\phi}{\partial u}+\frac{r_0^2}{p^2+r_0^2}\left[1+\frac{p^2}{(r_0+u)^2}\right]\frac{\partial^2\phi}{\partial w^2}=0\,. \qquad \ldots\ldots(29)$$

On the helical axis this becomes

$$\phi_{uu}+\phi_{ww}=-\frac{1}{r_0}\phi_0=\frac{2\phi_0}{p^2+r_0^2}, \qquad \ldots\ldots(30)$$

where we have made use of equations (18) and (11). Combining this with equation (25),

$$\phi_{uu}-\phi_{ww}=-2\phi_0\frac{p^2-3r_0^2}{(p^2+r_0^2)^2}, \qquad \ldots\ldots(25')$$

we obtain for the two derivatives of second order

$$\phi_{uu}=\frac{4\phi_0 r_0^2}{(p^2+r_0^2)^2}=-\frac{m}{e}\,\frac{2r_0^2}{p^2+r_0^2}\Omega^2, \qquad \ldots\ldots(31)$$

$$\phi_{ww}=\frac{2\phi_0(p^2-r_0^2)}{(p^2+r_0^2)^2}=-\frac{m}{e}\,\frac{p^2-r_0^2}{p^2+r_0^2}\Omega^2. \qquad \ldots\ldots(32)$$

Substituting these into (20), (21) we now obtain the equations of motion in an *anastigmatic system* in the form

$$\left.\begin{aligned}\ddot u&=-2\omega\dot w+\frac{p^2-r_0^2}{p^2+r_0^2}\Omega^2 u+\epsilon r_0\Omega^2,\\ \ddot w&=+2\omega\dot u+\frac{p^2-r_0^2}{p^2+r_0^2}\Omega^2 w.\end{aligned}\right\} \qquad \ldots\ldots(33)$$

§4. DISCUSSION OF THE ELECTRON-OPTICAL PROPERTIES

Apart from the term $\epsilon r_0\Omega^2$, equations (33) have the standard form of rotationally symmetrical systems. We can reduce them completely to this form by introducing, instead of u, the variable

$$u_\epsilon=u-u_0, \qquad \ldots\ldots(34)$$

where

$$u_0=-\epsilon r_0\frac{p^2+r_0^2}{p^2-r_0^2}, \qquad \ldots\ldots(35)$$

excluding the case $p=r_0$ which requires special treatment. This means that excess energy electrons perform their motion in the same way as standard electrons but around an axis $u=u_0$, which differs only radially from the helical axis. This is a direct generalization of the axis ' m ' in Figure 1 (*c*). Let us call it the ϵ axis. The equations are now

$$\ddot u_\epsilon=-2\omega\dot w-\alpha u_\epsilon, \qquad \ddot w=+2\omega\dot u_\epsilon-\alpha w. \qquad \ldots\ldots(36)$$

Multiplying the first by w, the second by u_e, subtracting and integrating as usual, we obtain the new momentum integral

$$u_\epsilon\dot w-w\dot u_\epsilon=\omega(u_\epsilon^2+w^2)+\text{const.} \qquad \ldots\ldots(37)$$

The constant is zero for any trajectory which crosses the ϵ axis. Such electrons therefore perform a *uniform rotation with angular velocity* ω around the axis. The reason is evident: it is our u, w coordinate system, which follows the electron which has an angular velocity component $-\omega$ in the direction of the helical axis.

The motion of such an electron which crosses the ϵ axis can be very simply described by introducing the new variable

$$\rho = (u_\epsilon^2 + w^2)^{1/2}, \qquad \ldots\ldots(38)$$

i.e. the distance from the ϵ axis. In terms of this, taking account of equation (37), the equation of motion becomes

$$\ddot{\rho} = -(\omega^2 + \alpha)\rho = -\frac{r_0^2}{p^2 + r_0^2}\Omega^2. \qquad \ldots\ldots(39)$$

This means that the electrons perform, apart from their rotation, a *radial simple harmonic oscillation* with a frequency

$$\frac{r_0}{(p^2 + r_0^2)^{1/2}}\Omega = \frac{r_0}{p}\omega. \qquad \ldots\ldots(40)$$

They return to the ϵ axis after a full period of this oscillation. But as they have the same energy, and as the length of the trajectories between the crossing points differs at most by quantities of the order ϵ^2, it follows that they will be in fact refocused in one point of the ϵ axis.

We are not, however, considering pencils of electrons, each focused on its own axis, but electrons of various energies, which have started from one point of the helical axis $u = w = 0$. The calculation, which need not be given in detail, shows that their motion is the resultant of two harmonic oscillations, with frequencies which are the combinations of the two simple frequencies ω and $(r_0/p)\omega$, previously considered, i.e. $(1 \pm r_0/p)\omega$. The full periods T_1 and T_2 of these two frequencies are given by

$$\Omega T_1 = 2\pi\frac{(p^2 + r_0^2)^{1/2}}{p + r_0}, \qquad \Omega T_2 = 2\pi\frac{(p^2 + r_0^2)^{1/2}}{p - r_0}. \qquad \ldots\ldots(41)$$

These formulae give directly the angles of rotation around the z-axis corresponding to these two periods.

If the two periods are incommensurable, the electron will not, in general, return to the helical axis. But if they are in an integral ratio, say n, they will recross it after the longer one, $T_2 = nT_1$, of the periods. It may be again noted that the time of return is independent of ϵ. But only monokinetic pencils will be focused at the same point, for reasons which we have explained in §2. The condition of zero '*chromatic astigmatism*' is

$$\frac{T_2}{T_1} = \frac{r_0 + p}{r_0 - p} = n = \text{integer}, \qquad \ldots\ldots(42)$$

which means that the pitch must be selected from the sequence

$$\tan\alpha = \frac{p}{r_0} = \frac{n-1}{n+1}. \qquad \ldots\ldots(43)$$

Suitable values are $\tan\alpha = 0, 1/3, 1/2, 3/5 \ldots$ The first has been already discussed in §2. In the other cases the electrons are refocused after 1·57, 2·23, 2·92 . . . full revolutions around the z axis.

These systems have no lateral chromatic error. The longitudinal error can be calculated in the same way as in §2. The excess energy electrons are moving on trajectories whose mean potential is lower by $\phi_u u_0$ than the potential ϕ_0 on the helical axis. As their initial energy is higher by a factor ϵ, their mean velocity excess is, in relative measure, $\frac{1}{2}(\epsilon + \phi_u u_0/\phi_0) = \frac{1}{2}\epsilon - r_0 u_0/(p^2 + r_0^2)$. On the other hand their trajectory is also longer by a fraction $r_0u_0/(p^2+r_0^2)$. Thus the focal shift is, using equation (35) for the value of u_0,

$$\tfrac{1}{2}\epsilon+\frac{2r_0u_0}{p^2+r_0^2}=\tfrac{1}{2}\epsilon+\frac{2r_0^2\epsilon}{p^2-r_0^2}=\tfrac{1}{2}\epsilon\frac{p^2+3r_0^2}{p^2-r_0^2} \qquad \ldots\ldots(44)$$

per unit length of trajectory. This function is illustrated in Figure 3. It starts for zero pitch with the value $-3/2$, which we have already obtained by an elementary consideration, and increases to infinity at $\tan\alpha=1$. At this point it changes sign, and decreases asympotically to the value $+\frac{1}{2}$. This too is easy to understand, as for very large pitch the curvature effects vanish, and the faster electron will be focused beyond the slower one, exactly in the ratio of its velocity.

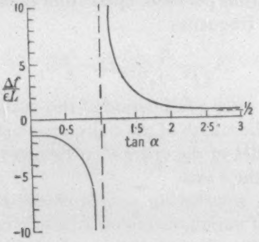


Figure 3. The chromatic shift of focus per unit increase of energy and unit length of trajectory.

§ 5. DETERMINING THE ELECTROSTATIC FIELD

The field in an anastigmatic helical system is not completely determined by the conditions (17), (18), (24) and (25) which we had to impose on it. Ultimately it may be perhaps possible to use the remaining freedom of design to eliminate the second-order, and perhaps even some of the third-order errors which we have left out of the present investigation. But at the present stage it will be best to make use of this freedom for simplifying the design as much as possible.

Laplace's equation (27) for the potential $\psi(r, z-p\theta)$ is

$$\frac{\partial^2\psi}{\partial r^2}+\frac{1}{r}\frac{\partial\psi}{\partial r}+\left(1+\frac{p^2}{r^2}\right)\frac{\partial^2\psi}{dz^2}=0. \qquad \ldots\ldots(45)$$

This equation gives the potential in a meridional plane, and allows the electrode shapes to be determined. Consider solutions of the form

$$\psi(r,z)=\frac{\cos}{\sin}\left(n\frac{z}{p}\right)R_n(r) \qquad \ldots\ldots(46)$$

for $\theta=0$, where n is an integer. The equation for R is

$$R_n''+\frac{1}{r}R_n'-\left(\frac{n^2}{p^2}+\frac{n^2}{r^2}\right)R_n=0, \qquad \ldots\ldots(47)$$

whose solution is

$$R_n=Z_n(jnr/p), \qquad \ldots\ldots(48)$$

where Z_n is a cylindrical function of order n. It is seen that an infinity of solutions exists which satisfies the conditions of first-order anastigmatism. But as we are looking for the simplest solution, we take $n=1$ only, and write the meridional field in the form

$$\psi(x,z)=\phi_0\left\{1-[ajJ_1(jx_0)+bH_1(jx_0)+\eta\log\frac{x}{x_0}+\cos\frac{z}{p}[ajJ_1(jx)+bH_1(jx)]\right\}. (49)$$

We have written here x for r/p and x_0 for r_0/p. The notations of Jahnke and Emde (1945) have been used for the cylindrical functions with imaginary argument. ψ assumes the value ϕ_0 on the helical axis $x=x_0$, $z=0$. It is symmetrical with respect to $z=0$, so that the conditions (17) and (24) are already satisfied. There remain only two conditions, (18) and (25) or (32). Equation (31) is already satisfied, as ψ is a solution of Laplace's equation. Using the identity

$$\frac{d}{dx}J_1(jx) = -\frac{1}{x}J_1(jx) + jJ_0(jx)$$

for the Bessel function J_1 and the corresponding identity for the Hankel function H_1, we now write the conditions (18) and (32) in the form : For $x=x_0$, $z=0$

$$-\frac{1}{\phi_0}\frac{\partial\psi}{\partial x} = -\frac{p}{\phi_0}\phi_u = -\frac{\eta}{x_0} + a\left(\frac{1}{x_0}jJ_1 + J_0\right) + b\left(\frac{1}{x_0}H_1 - jH_0\right) = \frac{2r_0p}{p^2+r_0^2}, \quad \text{..(50)}$$

$$-\frac{p^2}{\phi_0}\frac{\partial^2\psi}{\partial z^2} = -\frac{p^2r_0^2}{p^2+r_0^2}\frac{1}{\phi_0}\phi_{ww} = ajJ_1 + bH_1 = -\frac{2p^2r_0^2(p^2-r_0^2)}{(p^2+r_0^2)^3}. \quad \text{......(51)}$$

With these conditions we have disposed of only two of the three parameters a, b and η. We use the remaining degree of freedom to simplify the design, by imposing the condition that the equipotential surface at a radius r_i shall be cylindrical, i.e. for $x=x_i$

$$ajJ_1(jx_i) + bH_1(jx_i) = 0. \quad \text{......(52)}$$

An example of the electrode design obtained from these equations * is shown in Figure 4. The pitch is one of the series (43) for which the chromatic astigmatism is zero. The radius r_i of the inner, cylindrical electrode is made rather small, one-quarter of the radius of the helical axis. This is advantageous as it produces a smooth field outside r_0; moreover, in all practical cases one would connect the inner electrode with the highest positive potential available, and a large ratio r_0/r_i means a relatively low potential on the helical axis. This in turn means that a relative energy difference of 1% produces in our case an ϵ of 3·55%, so that the defocusing is not 2% as appears from the curve in Figure 3, but about 7% of the length of the helical axis. There is of course a limit to the slowing down of electrons in the helical systems; in the end one would incur the danger of uncontrollable errors by polarization layers on the electrodes, which is one of the chief objections against mirrors with straight axis.

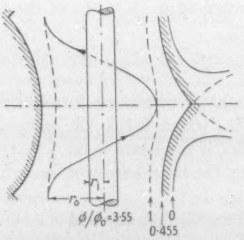


Figure 4. Electrode design for a helical lens, $p/r_0 = \tan\alpha = 0{\cdot}5$, $r_i/r_0 = 0{\cdot}25$.

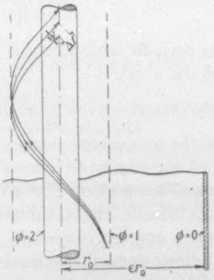


Figure 5. Helical velocity spectrograph, $\alpha = 45°$.

---

* Computed by Mr. K. J. Ausburn.

It is advantageous to select for the outer electrode the equipotential line which passes through the saddle point, where the field is *nil*, as small deformations of this line have minimum influence on the field at the helical axis; also the sharp corner makes accurate mounting easier.

The design as shown is of course not complete in itself, one must add tubes at both ends for introducing and removing the electron beam. These tubes necessarily modify the field at both ends somewhat, and it is intended first to study these effects experimentally before proceeding further with the theory. On the other hand it may be possible to compensate some aberrations by suitable injecting and extracting organs; in particular it may be possible to add deflectors which allow alignment of the initial and the final tangent, whose skew relative position is a considerable inconvenience in the present design.

§6. A VELOCITY SPECTROGRAPH

The critical case $p = r_0$, i.e. $\tan\alpha = 1$, requires special examination as here the method used for integrating the equations (33) breaks down. The equations now become

$$\ddot{u} + 2\omega\dot{w} = \epsilon r_0 \Omega^2; \quad \ddot{w} - 2\omega\dot{u} = 0. \qquad \ldots\ldots(53)$$

A particular integral of the inhomogeneous equations is

$$u = 0; \quad w = \frac{\epsilon r_0}{2\omega}\Omega^2 t. \qquad \ldots\ldots(54)$$

This means that while in all other cases the 'ϵ axis' differed from the helical axis only radially, in the direction of u, and had the same pitch, now it differs in the direction of the binormal w, and has a different pitch. The complete solution can be written in the form

$$u = C_1 e^{2j\omega t}, \qquad w = C_2 e^{2j\omega t} + \frac{\epsilon r_0}{2\omega}\Omega^2 t. \qquad \ldots\ldots(55)$$

This is periodic with a period T, which we write again in terms of the rotation around the z axis:

$$T = \frac{\pi}{\omega} = \pi\frac{(p^2 + r_0^2)^{1/2}}{p} = \sqrt{2}\pi, \qquad \ldots\ldots(56)$$

that is, the monokinetic beams issuing from a point come to a focus in the time in which the standard electron rotates by $\sqrt{2}\pi$, which is just the same angle as in the well-known case of two coaxial cylinders, if the beam is launched at right angles to the axis. But the difference now is that the beams are *completely* focused, and not in one direction only.

It follows immediately from equations (50) to (52) that the field is a cylindrical one, with a law

$$\phi = \phi_0(1 - \log r/r_0), \qquad \ldots\ldots(57)$$

which can be produced by the electrode arrangement shown in Figure 5. Thus this system is of the greatest simplicity. Moreover, it has very useful properties as a velocity spectrograph, as the beams of different energy are all focused in a meridian plane. This can be easily verified. By equation (54) the ϵ axis diverges from the helical axis by $w = \frac{1}{2}\epsilon r_0\,(\Omega/\omega)\Omega T = \pi\epsilon r_0$. The ϵ axes are now all on the same cylindrical equipotential surface as the helical axis, hence in the time T the faster electron runs past the standard electron by $\frac{1}{2}\epsilon \times$ the length of the trajectory or, as this length is $2\pi r_0$, by $\Delta f=\pi\epsilon r_0=w$. One can also verify that on the ϵ axes as shown in Figure 5 the tangential velocity of all electrons is the same; they differ only in their z component, hence they perform the same rotation around the axis in the time T.

The separation of the foci of electrons differing by an energy fraction ϵ is therefore $z=\sqrt{2\pi\epsilon r_0}$, and this can be increased to a considerable extent by slowing down the electrons, as far as it is safe to go.

It is probable that interesting variations of this spectrograph could be obtained by combining the radial electrostatic field with an axial magnetic field.

REFERENCES

COTTE, M., 1938, *Thèse* (Paris: Masson et Cie).

GABOR, D., 1944, *The Electron Microscope* (London: Electronic Engineering); 1949, *Proc. Roy. Soc.* A, **197**, 454.

HUTTER, R. G. E., *Advances in Electronics*, edited by L. Marton (New York: Academic Press), p. 167.

JAHNKE, F., and EMDE, E., 1945, *Tables of Functions with Formulae and Curves* (New York: Dover Publications), pp. 228, 236.

RAMBERG, E. G., 1949, *J. Appl. Phys.*, **20**, 183.

SCHERZER, O., 1936, *Z. Phys.*, **101**, 593.

WENDT, G., 1943, *Z. Phys.*, **120**, 720.

An Investigation of certain Properties of the Ionosphere by Means of a Rapid Frequency-Change Experiment

BY B. H. BRIGGS

Cavendish Laboratory, Cambridge

Communicated by W. L. Bragg; MS. received 25th October 1950

ABSTRACT. The paper is concerned with the recording of various characteristics of a radio wave reflected from the ionosphere as the frequency is altered rapidly.

An apparatus is described in which a receiver is kept in tune with a sender by means of automatic frequency control circuits operated by the direct pulse from the sender. The tuning is corrected by each pulse and remains fixed for the time between pulses. A frequency range of 1 Mc/s. can be covered in 1 sec. and may be selected anywhere in the range 2–20 Mc/s.

A number of applications of the apparatus are described and typical records are reproduced. The irregularities of ionization present in the normal regions are investigated by the observation of the irregular variations of the amplitude of the reflected wave which are produced when the frequency is altered. Irregularities of ionization are sometimes found to be localized in height. Observations of the reflection coefficient of the abnormal region E suggest that there are two distinct types of region, one an irregular region consisting of scattering clouds, the other a coherent layer with a thickness of the order of 5 km. The apparatus is also used to study the behaviour of the subsidiary critical frequencies often present below the main critical frequency of region E. It is found that these are always decreasing in critical frequency whenever they are observed. Records of amplitude and group path near the critical frequency of a region can be used to determine the collision frequencies of electrons. The treatment is mainly descriptive, and a full discussion of those records which require a detailed quantitative analysis is reserved for a later paper.

§ 1. INTRODUCTION

In the frequency range 1–20 Mc/s. one of the most useful methods of ionospheric investigation is the pulse method first used by Breit and Tuve (1926). Measurements of the delay time of the echo give the group path P', and attention has in the past been directed to the recording of the variations of P' as the frequency is changed. Several descriptions of apparatus designed to record (P',f) curves automatically have already appeared.

In the present paper we are mainly concerned with the variations of the *amplitude* R of the reflected wave as the frequency is changed. We shall refer to a record of these variations as an (R, f) record*. The design of apparatus to produce (R, f) records is more difficult than in the case of (P', f) records for two reasons. Firstly, the measured amplitude will not be correct unless the receiver is exactly in tune with the sender when the echo arrives. The value of P', on the other hand, is not critically dependent on the tuning. Secondly, it is necessary to obtain the complete (R, f) record very quickly if the results are to be significant, because the value of R on each frequency is found to vary in an irregular way with time (i.e. the echo is subject to 'fading'). In the recording of (P', f) curves high-speed recording is also desirable because the value of P' may change with time, but in general the changes of P' take place much more slowly than the changes of R.

The design of an apparatus which combines high speed of operation with accuracy of tuning of the receiver is described in § 2. This apparatus is suitable for obtaining (R,f) records. It can, of course, also be used to obtain (P',f) records if required. For the applications described in this paper it is sufficient, and indeed desirable, to confine the observations at any one time to a relatively small band of frequencies of the order of 1 Mc/s. or less in width. The apparatus is capable of producing records for a frequency band of this order of width in a time of 1 second. The frequency band in use may be selected to lie anywhere in the range of 2–20 Mc/s.

The remainder of the paper is concerned with a discussion of the different types of record which can be obtained with the equipment. Some theoretical ideas which are needed in the discussion of the records are first outlined in § 3. Then in §§ 4 and 5 we take up the discussion of the records, the main topics being (*a*) the irregularities of the structure of regions E and F, (*b*) the nature of the abnormal region E, and (*c*) the behaviour of the subsidiary critical frequencies associated with region E. These phenomena are investigated by the use of two different types of record; § 4 deals with single (R, f) and (P', f) records, and § 5 with 'panoramic' records, which enable amplitude changes to be observed as a function of time as well as of frequency. Another application of the apparatus is to the determination of the collision frequencies of electrons at various heights in the ionosphere, and a separate paper will be written on this subject. The method is given in outline in § 6.

§ 2. DESCRIPTION OF THE APPARATUS

The equipment consists essentially of a pulse sender, provided with a wide-band aerial system, and a receiver, also having a wide-band aerial. The frequency of the sender is varied cyclically backwards and forwards over a frequency band, the variation being approximately linear, and at the same rate for both increasing and decreasing frequency. The receiver is provided with automatic frequency control circuits which keep it in tune with the sender. The variations of the echoes returned from the ionosphere are examined and recorded in various ways to be described later.

---

* The more usual notation (ρ, f)—reflection coefficient ρ as a function of frequency—has been avoided as a general description of the records. This notation will be used rather to describe the result of averaging a number of (R, f) records, so that fluctuations due to random diffraction effects are eliminated (§ 6).

Some parts of the apparatus follow standard practice and will not be described in any detail. The main features of interest are the method of automatic frequency control used in the receiver and the arrangements used to display the results.

(i) *The Sender and the Aerial Systems*

The sender uses a pair of 813 valves in a conventional self-oscillator circuit. Radio-frequency pulses are produced by the application of a positive modulation pulse to the screen grids of the valves. The peak power developed is about 5 kw., and the pulse length 200 μsec. The frequency sweep is produced by a simple mechanical movement of the main tuning condenser, and provision is made for adjustment of the extent of the sweep and the centre frequency.

The pulse used to modulate the sender is also passed to the receiver, where it is used to trigger various circuits as described later. The sender and the receiver are situated close together in the same hut. Wide-band aerials are necessary both for sending and receiving. The aerial used in conjunction with the sender consists of an inverted 'V' terminated at the top with a 900 Ω resistor, and fed with an open wire line. The receiving aerial system consists of a pair of equal dipoles at right angles. The bandwidth of the dipoles is increased by constructing each half with three wires, instead of the usual single one, the wires being arranged in the shape of a fan. The overall length of each dipole is 85 feet.

The object of using a *pair* of crossed dipoles is to permit the selection of either a right-handed or a left-handed circularly polarized downcoming wave. This can be done if the two outputs are added together after the introduction of a relative phase shift of 90°. In the present apparatus, circuits are used which produce the required 90° phase shift over a wide frequency range without the necessity for adjustment (Phillips 1951). By reversing the connections to one aerial either of the two circularly polarized components may be selected.

The use of an aerial system of this type is essential in some of the experiments, because it is necessary to observe a single magneto-ionic component, so that fluctuations of amplitude which arise from the interference between the two magneto-ionic components are eliminated.

When a simple linearly polarized aerial is required, one dipole alone is used.

(ii) *The Receiving Apparatus*

The receiver has an overall bandwidth of 20 kc/s., and must therefore be kept in tune with the sender to an accuracy of a few kc/s. At the maximum rate of sweep the frequency of the sender will be varying over a range of 1 Mc/s. in one second; as the pulse repetition rate is 50 per second, the frequency difference between successive pulses is then 20 kc/s. Thus the individual pulses find the receiver completely off tune. The frequency control must therefore operate during the time occupied by the direct pulse from the sender. Also the tuning must remain fixed for $\frac{1}{50}$ second after each pulse has ended, so that the receiver will be tuned to the frequency of the echoes when they arrive.

The design of automatic frequency control circuits for pulse operation has been discussed by Moxon, Croney, Johnston and Laws (1946). They point out, however, that all such systems used in the past have been of a 'slow-acting' type, a correcting voltage being obtained by averaging over a number of pulses. Owing to the large frequency difference between successive pulses, such a system would not be capable of correcting the tuning with sufficient accuracy in the present case. It was therefore necessary to develop some new circuits which would be very rapid in operation, and which would maintain correct tuning of the receiver in between the controlling pulses.

A block diagram of the receiving system is shown in Figure 1. The incoming signals pass first through a band-pass high-frequency amplifier which has a gain of about 30 times. The bandwidth is approximately 1 Mc/s. and the amplifier is tuned manually so that this band covers the same frequency range as that in use at the sender. The signals from the amplifier pass to Mixer I, where they are converted to an intermediate frequency of 20 Mc/s. The oscillator feeding Mixer I is variable in frequency from 22–40 Mc/s., and is set manually to approximately the correct frequency to give the 20 Mc/s. intermediate frequency, depending on the centre frequency of the incoming signals, which may be between 2 and 20 Mc/s. The frequency is then automatically controlled, as described in detail below, so as to maintain the oscillator frequency at the value which gives exactly 20 Mc/s. output from Mixer I, even when the frequency of the sender is varying. The high intermediate frequency of 20 Mc/s. is used so that the frequency control may be applied to an oscillator the frequency deviation of which is only a small fraction of its actual frequency.

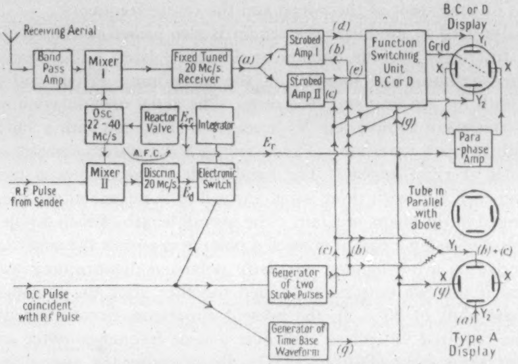

Receiving Aerial
Band Pass Amp
Mixer I
Fixed Tuned 20 Mc/s. Receiver
(a)
Strobed Amp I
Strobed Amp II
(d)
(b)
(e)
(c)
Function Switching Unit B, C or D
B, C or D Display
Grid
Y_1
X
X
Y_2
Osc 22 - 40 Mc/s
Reactor Valve
E_r
Integrator
A.F.C.
E_r
(g)
Para-phase Amp
Mixer II
Discrim. 20 Mc/s.
E_d
Electronic Switch
R.F. Pulse from Sender
Tube in Parallel with above
(c)
(b)
Y_1
(b) + (c)
Generator of two Strobe Pulses
X
(g)
X
D.C. Pulse coincident with R.F. Pulse
(a)
Y_2
Generator of Time Base Waveform
(g)
Type A Display

Figure 1. Block diagram of the receiving apparatus.

The output of Mixer I passes to a receiver tuned to a fixed frequency of 20 Mc/s. where the main amplification and selectivity are obtained. The rectified output of this receiver then passes to the various display systems and strobe selectors, which are discussed in (iv) below.

The action of the frequency control will now be described. It will be seen in Figure 1 that the oscillator feeds Mixer II as well as Mixer I. This second mixer receives a strong ground pulse direct from the sender (Figure 2 (*a*)), and produces a pulse at the difference frequency between the sender and the oscillator. This difference frequency passes to a discriminator, from which is obtained a positive or a negative pulse, depending on whether the difference frequency is above or below 20 Mc/s. The pulse from the discriminator passes through an electronic switch to an integrator and then to a reactor valve which controls the frequency of the oscillator. When the frequency control is operating correctly, the discriminator output E_d will be as shown in Figure 2 (*b*). There will be an initial output from the discriminator when each pulse arrives from the sender, but this will fall to zero during the pulse, showing that the frequency error has been corrected. The output of the integrator, E_r, will be as shown in Figure 2 (*c*); this waveform shows how the tuning is maintained at the correct value during the intervals between the pulses. The same waveforms are drawn with a compressed time scale in Figures 2 (*e*) and (*f*), in order to show one complete cycle of the frequency variation. A typical variation of the frequency of the sender with time is shown in Figure 2 (*d*) for comparison with the voltage waveforms existing in the receiver. Since E_r is the voltage applied to the reactor valve, it is proportional to the deviation of the frequency of the sender, the reactor valve being very nearly linear in response. The electronic switch between the discriminator and the integrator serves two purposes. Firstly, as the switch is closed only during the time occupied by the pulses, the frequency control is inoperative except during these periods, and so the system cannot be affected by interfering signals. Secondly, the presence of the switch simplifies the design of the integrator, which can, in fact, simply be a condenser. No current can flow into this condenser in the intervals between pulses, because it is completely disconnected from the discriminator.

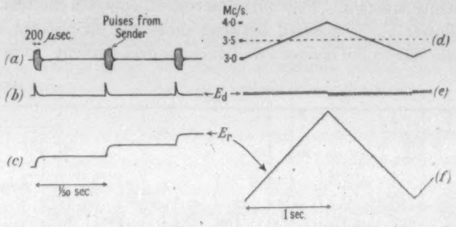


Figure 2. (*a*) Radio-frequency pulses from sender; (*b*) and (*c*) waveforms of E_d and E_r (time intervals not to scale); (*d*) typical variations of the frequency of the sender; (*e*) and (*f*) waveforms as in (*b*) and (*c*) showing one complete cycle of the frequency variation.

(iii) *The Automatic Frequency Control Circuits*

In general we shall not discuss the detailed circuits of the various units shown in the block diagram of Figure 1. The circuits of those units which produce the frequency control action are, however, of some interest, and will be described in this section.

The relevant circuit is shown in Figure 3. The circuit has been simplified by the omission of decoupling components and other details. The first valve V_1 is Mixer II of Figure 1. It receives the direct radio-frequency pulse from the sender and also the output of the oscillator V_{13}, and produces the difference frequency of 20 Mc/s. in its anode circuit. The output of V_1 passes to the discriminator valves V_2 and V_3. The tuned circuits in the anodes of these valves are tuned so that one is resonant slightly below 20 Mc/s. and the other slightly above 20 Mc/s. The outputs of the two tuned circuits are rectified by V_4 and V_5 and the difference of the two rectified outputs passes to the amplifier V_6. This type of discriminator was found to be excellent for pulse working because of its perfect symmetry. The output pulse from the amplifier V_6 may be either positive or negative, depending on whether the frequency of the sender is increasing or decreasing. This pulse passes through the electronic switch formed by the four diodes V_7–V_{10}, and either charges or discharges the integrating condenser C_1. The action of the diode switch is as follows: In the interval between pulses, the diodes V_8 and V_{10} are conducting, and V_7 and V_9 are non-conducting. Thus no current can flow into the condenser C_1 from V_6. During the time occupied by the pulses from the sender, controlling pulses are applied to the diodes V_8 and V_{10} so as to render them non-conducting. A positive pulse from V_6 can now reach condenser C_1 via the resistor of 150 kΩ and the diode V_9; similarly, a negative pulse can reach the condenser C_1 via V_7. Thus the output of V_6 is effectively connected to C_1 via a resistance of 150 kΩ for the duration of the controlling pulses. In between pulses C_1 is disconnected from the rest of the circuit, and its charge remains constant. The voltage across C_1 controls the reactor valve V_{12}. A cathode follower V_{11} is inserted between these two to prevent any discharge of the condenser due to the reactor valve drawing grid current. The reactor valve in turn controls the frequency of the oscillator V_{13}, the output of which is fed to V_1, thus completing the automatic frequency control loop.

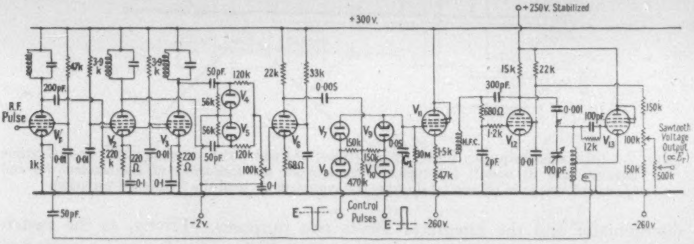

+300 V.
+250 V. Stabilized
R.F. Pulse
V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
V11
V12
V13
C1
200pF.
47k
3·9 k
3·9 k
1k
0·01
0·01
220 k
220 Ω
0·01
220 Ω
0·1
0·1
50pF.
120k
56k
56k
50pF.
120k
100k
0·1
22k
33k
0·005
68Ω
150k
150k
470k
0·05
30M
1·5k
47k
H.F.C.
15k
22k
300pF.
680Ω
1·2k
2pF.
0·01
0·001
100pF.
100pF.
12k
150k
100k
150k
500k
Sawtooth Voltage Output (∝ E_r)
50pF.
-2v.
E
Control Pulses
E
-260v.
-260v

Figure 3. Circuit diagram of the sections of the receiving apparatus which produce the automatic frequency control.

The response of this circuit will now be considered in order to show that the required results are attainable. The valves V_1–V_6 can be classed together as a 'discriminator unit'. This unit has the property that an error of 1 Mc/s. in the frequency applied to V_1 results in an output of A volts, say, at the anode of V_6. This voltage does not appear instantaneously, because of the build-up time of the tuned circuits associated with the valves V_1, V_2 and V_3. Let this build-up be represented by a time constant T_1. Next consider the 'reactor valve unit' comprising V_{11}, V_{12} and V_{13}. This has the property that B volts, say, applied to the grid of V_{11} produces a change of oscillator frequency of 1 Mc/s. There is no time lag present here. Let the main integrator associated with the diode switch have a time constant T_2. It will be seen that the loop can be regarded as having an overall 'amplification' of A/B, and it contains two time constants T_1 and T_2. It can be shown that the condition for a critically damped response is

$$A/B = \tfrac{1}{4}T_2/T_1, \qquad \ldots\ldots(1)$$

provided $T_2 \gg T_1$ and $A/B \gg 1$. (Both these conditions are satisfied in practice.) Since $T_2 \gg T_1$ the main time scale of the response can be obtained by ignoring the smaller time constant T_1. In this case the response becomes an exponential approach towards the correct tuning point, the time constant of the exponential being $T_2 B/A$. Now the frequency difference between successive pulses is 20 kc/s., and if a tuning error of 2 kc/s. is tolerated, this means that the error must be reduced to at least one-tenth of its initial value during the time occupied by each pulse. This requires a time equal to approximately two time constants. Thus if T is the time occupied by each pulse, we must have

$$2T_2 B/A \leqslant T. \qquad \ldots\ldots(2)$$

Combining equations (1) and (2) the condition which must be satisfied is found to be

$$T_1 \leqslant T/8. \qquad \ldots\ldots(3)$$

Now the duration of the pulses from the sender is normally about 200 μsec.; consequently T_1, the build-up time of the tuned circuits of the discriminator, must be less than 25 μsec. This condition is not difficult to satisfy.

Great care has been taken to avoid the introduction of any additional time constants in the feedback loop. It is difficult to predict the effect of the small stray capacities which must be present at various points, but in practice no difficulty was experienced in securing satisfactory operation.

(iv) *Cathode-Ray Tube Displays, and Recording Arrangements*

In this section the remaining units of Figure 1 will be described. These are concerned with the methods used for displaying and recording the results. Figure 4 shows the various waveforms which are involved (*a*) to (*g*), and the four types of display (A to D). Lettering (*a*) to (*g*) corresponds with that used in Figure 1.

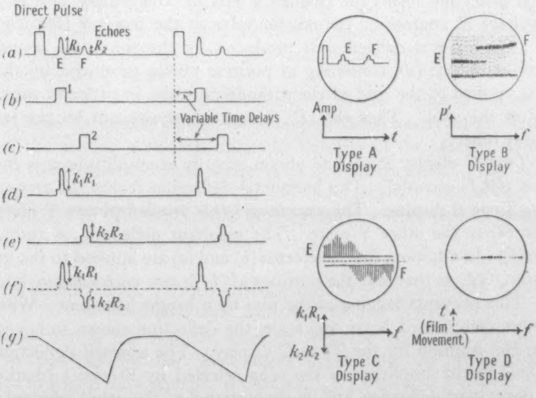


Figure 4. Waveforms (*a*) to (*g*) and display systems (A to D).

For recording purposes it is necessary to isolate a particular echo or echoes from the ground pulse and other echoes which may be present. The ***strobe generator*** unit of Figure 1 produces two ***strobe pulses*** (Nos. 1 and 2) at adjustable times after the pulse from the sender. The waveforms produced by this unit are shown in (*b*) and (*c*) of Figure 4. Let us suppose that there are present echoes from regions E and F, so that the waveform at the output of the receiver is as shown in (*a*) of Figure 4, and that we wish to record the fluctuations of amplitude of both the echoes. We should then adjust the positions of the two strobes so that No. 1 embraces the echo from region E and No. 2 embraces the echo from region F, as shown in the diagram. The signals from the receiver pass to two ***strobed amplifiers***, one controlled by each strobe pulse. Such an amplifier is one which operates only during the period occupied by the strobe pulse; for the remainder of the time no signal can reach the output. Thus the waveforms at the outputs of the strobed amplifiers appear as shown in (*d*) and (*e*) of Figure 4. Each echo appears on top of a 'pedestal', because a certain fraction of each strobe pulse is allowed to reach the output.

The generator of time-base waveform shown in Figure 1 produces the time-base waveform (*g*) of Figure 4; it is triggered by the D.C. pulse from the sender.

The methods by which the various waveforms shown in Figure 4 are produced have now been described. We shall consider next how these waveforms are used to produce the various types of display.

The Type A display is simply a range display of a conventional type. The horizontal deflection of the cathode-ray tube consists of the time-base waveform (*g*), and the vertical deflection is produced by the receiver output (*a*), and by the waveforms (*b*) and (*c*). In this way the positions of the strobe pulses are shown. The Type A display is not used for recording purposes, but is used to keep the positions of the echoes and the strobes under continual observation.

The Type B display is used when it is desired to record (P',f) curves. The horizontal deflection forms the frequency axis for the record, and is produced by the voltage E_r controlling the reactor valve in the receiver (see Figures 1, 2 and 3). The vertical deflection is produced by the time-base waveform (*g*). The receiver output (*a*), consisting of positive pulses produced by the various echoes, is applied to the grid of the cathode-ray tube in order to modulate the intensity of the spot. Thus the (P',f) curve is drawn out on the face of the cathode-ray tube.

The Type C display is used to obtain records of amplitude as a function of frequency ((R,f) records). The horizontal deflection forms the frequency axis as for the Type B display. The waveform (*d*) is applied to one Y plate and the waveform (*e*) to the other Y plate. The resultant deflection of the spot is as shown in (*f*). In addition, the waveforms (*b*) and (*c*) are applied to the grid of the cathode-ray tube, so that only the portions of (*f*) drawn with a heavy line actually appear. This prevents fogging of the film by a bright base-line. When drawn with a more compressed horizontal scale, the deflection shown in (*f*) appears as shown in the diagram for the Type C display. The upward deflection k_1R_1 is proportional to the amplitude of the echo selected by Strobe 1 (marked as R_1) and the downward deflection k_2R_2 is proportional to the other selected echo R_2. Thus the amplitudes of both are recorded simultaneously as functions of frequency. The constants k_1, k_2 depend upon the amplification used in the two strobed amplifiers. When observing two echoes which differ considerably in strength, it is convenient to have k_1 and k_2 unequal, and the two amplifiers are therefore provided with separate gain controls. As will be clear from the waveform (*f*), the zero lines for the two records are not coincident, but are separated by a certain distance. For examples of the records, see Figures 6 to 11 (Plate I)*.

The Type D display is used when it is desired to obtain a continuous record of the amplitude of a single echo over a certain frequency band as a function of time. The horizontal deflection again represents frequency, as in B and C; no vertical deflection is used. The waveform (*d*) is applied to the grid of the cathode-ray tube. Thus the spot is bright for those frequencies and times when the signal is strong, and is weak when the signal is weak. A film moves slowly past the face of the cathode-ray tube in a vertical direction and the required information is recorded as an intensity modulation on the film. The recording of amplitude is not quantitative, but nevertheless the records are found to be very valuable in the investigation of a number of phenomena. Records of this type will be called *panoramic records*; examples are shown in Figures 13 to 24 (Plates II–IV).

With the use of a sufficient number of cathode-ray tubes all the types of display could be in operation simultaneously. It was decided, however, to make the three types of display B, C and D, alternatives, the type required being selected by a switch. The 'function switching unit' of Figure 1 accepts the various waveforms and connects them in the appropriate ways to the electrodes of two cathode-ray tubes in parallel. One of the cathode-ray tubes is used for visual observations, and the other for photographic recordings. The tube used for visual observations is of the long-afterglow type. The (P',f) and (R,f) records consist of photographs of single frequency sweeps, and as the time occupied by one sweep is normally 1 second, the amplitudes of 50 pulses are recorded. This is usually sufficient to give adequate resolution of the fluctuations. To produce the panoramic records, the Type D display is used in conjunction with the continuous movement of a film, and the camera shutter is left open. It is found to be advantageous to double the rate of sweep for this type of recording, so that each sweep occupies $\frac{1}{2}$ second.

It is interesting to note that the linearity of the frequency scale in the Type B, C, and D displays depends not upon the linearity of the frequency variation of the sender itself, but upon the linearity of the reactor valve in the receiver. If the frequency of the sender does not change linearly, the only effect is to produce an unequal spacing of the individual pulses on the records. Thus it is not necessary to take any special precautions to ensure an accurately linear variation of the frequency of the sender.

§ 3. THEORETICAL IDEAS CONCERNING REFLECTION FROM AN IRREGULAR REFLECTING LAYER

Before discussing the records obtained with the apparatus, we shall first consider in this section the basic ideas concerning the interpretation of experiments in which the frequency is changed. We shall also discuss in a simple, semi-quantitative manner the theoretical treatment of reflection from an irregular reflecting layer, and the theory of fading of ionospheric echoes.

Consider first a situation in which there are two paths by which a signal can travel from the sender to the receiver. We then expect the amplitude to vary periodically as the frequency is changed, the periodicity being determined by the value of the path difference. This effect was used by Appleton and

---

* For Plates see end of issue.

Barnett in their first proof of the existence of a reflecting layer in the upper atmosphere. The two paths were provided by the ground wave and the wave reflected from the layer. Appleton showed that the results of such experiments depend on the difference of the *group paths* of the two waves. Thus, if the group paths are P_1' and P_2', and if Δn maxima or minima are observed in a frequency range Δf we have

$$\Delta n/\Delta f=(P_1'-P_2')/\mathbf{c}. \qquad \ldots\ldots(4)$$

In the experiments with which we are concerned in the present paper, interference between the ground wave and the sky wave, and between multiply-reflected waves is eliminated by the use of pulses. Nevertheless, it is frequently found that fluctuations of amplitude occur as the frequency is varied. These changes are usually irregular rather than periodic, and must be due to the fact that the ionosphere is not a specular reflector but is irregular.

Reflection from an irregular ionosphere has been studied theoretically by Booker, Ratcliffe and Shinn (1950), and experimentally by Briggs and Phillips (1950). It has been shown that the downcoming wave consists of a cone of rays spread over a certain range of angles $\pm\theta_0$, as shown in Figure 5, the value of θ_0 being different on different occasions. The path differences which account for the irregular changes of amplitude as the frequency is varied will arise mainly from the obliquity of the various rays in the cone. A continuous distribution of such path differences will be present; this accounts for the random nature of the variations. If in an (R,f) record we count the number of maxima Δn occurring in a frequency range Δf, we can obtain an estimate of the order of magnitude of the *maximum* path difference which is present. Now from Figure 5 it is easily shown that this maximum path difference is $2\delta h=2h(\sec\theta_0-1)$, where h is the height of the layer. Consequently, the following equation should hold approximately:

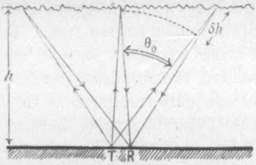


Figure 5. Reflection at an irregular ionosphere.

$$\Delta n/\Delta f=(2h/\mathbf{c})\,(\sec\theta_0-1). \qquad \ldots\ldots(5)$$

This result will be required in the discussion of (R,f) records in §4.

The experimental values of θ_0 given by Briggs and Phillips (1950) were obtained by observing the differences in the fading at spaced receiving points. The results of such experiments depend upon the structure of the irregular diffraction pattern produced on the ground, which is related by the theory to the degree of angular spreading of the downcoming wave. It is now clear that a frequency-change experiment provides an alternative method of studying the properties of an irregular reflecting surface. The theory just given relates the two methods, and it is valid so long as the reflection process is confined to a well-defined horizontal layer as pictured in Figure 5. It is possible that under some ionospheric conditions there may be more complicated effects such as scattering in depth, or reflection from one layer and scattering from a different layer beneath, and there will then be no simple relationship between the results of the two experiments.

It is possible to put the argument leading to equation (5) on a more quantitative basis by the use of an auto-correlation function to describe the changes of amplitude as the frequency is varied. This function can be related to the distribution function which gives the fraction of the energy which travels different distances in passing from the sender to the receiver. As we do not intend to deal with the full statistical analysis of the records, we shall not develop these ideas further here. The semi-quantitative argument given above will be sufficient for the purposes of the present paper.

When discussing panoramic records in §5 we shall be concerned not only with the properties of an (R,f) record at a particular instant, but also with the changes in *time*. It is therefore necessary to outline the present theories of the fading of ionospheric echoes. It has been pointed out already that reflection at an irregular region gives rise to a spreading of the downcoming wave in angle, and to an irregular diffraction pattern of amplitude over the ground. If the configuration of the ionosphere did not change, the amplitude observed at fixed points on the ground would not change with time. Changes in the pattern must be due to movements in the ionosphere. They may be of two types: (*a*) a steady drift of the pattern past the observing points due to a horizontal movement of the ionosphere, and (*b*) random changes, due to random movements of different parts of the ionosphere. Briggs, Phillips and Shinn (1950) have discussed these two causes of fading, and have shown that they may be separated by a suitable analysis of the fading records obtained at three points on the ground. From the results of such experiments it is believed that a horizontal movement of the reflecting layer is the main cause of fading, the effect of random changes being less important. If we accept this picture, we can derive an approximate expression for the number of maxima occurring per unit time in a record of the fading. In Figure 5 we represented the downcoming wave as a cone of rays spread over a range of angles $\pm\theta_0$. Let the ionosphere be given a horizontal motion of velocity V. The rays in the cone which travel at an angle θ_0 to the vertical suffer on reflection a Doppler frequency shift $\pm(2V/\lambda)\sin\theta_0$. The frequency difference between the extreme rays in the cone is $(4V/\lambda)\sin\theta_0$, and this expression will give the highest frequency present in the fading record. Thus, if Δn maxima occur in a time Δt, the following result will be approximately true

$$\Delta n/\Delta t = (4V/\lambda)\sin\theta_0. \qquad \ldots\ldots(6)$$

This equation will be needed in the interpretation of some of the records discussed in §5.

§4. EXPERIMENTAL RESULTS: RECORDS OF AMPLITUDE AND GROUP PATH AS FUNCTIONS OF FREQUENCY

In this Section some records of the (R,f) and (P',f) types will be discussed. Typical records are shown in Figures 6 to 11.

(i) *Records showing Irregular Fluctuations of Amplitude*

Consider first Figures 6 and 7 which are records of amplitude alone (Type C display). The frequency range was well away from any critical frequencies. The upper record (1F) in each case represents the strength of the signal which has suffered a single reflection at region F, and the lower record (2F) shows the strength of the signal which has suffered two reflections at region F with an intermediate reflection at the ground. In order to make the 2F echo comparable in amplitude to the 1F echo on the record, more amplification was used in the strobed amplifier which selected the 2F echo than in the amplifier which selected the 1F echo.

It will be observed that, in Figure 6, both the 1F and the 2F echoes show irregular fluctuations of amplitude. These fluctuations must be caused by reflection at a 'rough' ionospheric region. The number of maxima observed within a given frequency band is found to vary considerably on different occasions. Figure 7 is an example in which there is hardly any fluctuation of the amplitude of the 1F echo. This illustrates a rather rare condition in which the ionosphere behaves as a specular reflector. The 2F echo still shows some fluctuations, though less marked than in the previous example. These fluctuations must be connected with irregular reflection at the ground.

The record for the 1F reflection shown in Figure 6 shows eight maxima in a frequency range of 400 kc/s. and the height of the layer was 325 km. Hence, from equation (5) the value of θ_0, giving the degree of angular spreading of the downcoming wave, is found to be 7·2°. In general, the values of θ_0 determined in this way are in agreement with the values obtained from observations of fading at spaced receiving points (Briggs and Phillips 1950).

(ii) *Critical Frequency Phenomena for the Normal Ionospheric Regions*

Consider next phenomena occurring near the critical frequencies of the two main ionospheric regions. Figures 8 and 9 show examples of such phenomena. In each case a curve of equivalent height ($\frac{1}{2}P'$) against frequency is shown (Type B display) together with the corresponding records of amplitude (Type C display).

Figure 8 shows the penetration of region F as the critical frequency was passed. The upper curve is the well known (P', f) curve, and the lower record shows how the amplitude of the echo decreased as the critical frequency was approached.

Figure 9 illustrates the more complex effects sometimes observed near the critical frequency of region E. The (P', f) curve in the upper part of the Figure indicates that there are multiple penetration effects, so that the curve of ionization against height would, on this occasion, have a number of subsidiary maxima at various heights between 100 km. and 150 km. In the lower part of Figure 9 the amplitude of the echoes returned from region E are shown recorded upwards, and the amplitude of the echo from region F is recorded downwards. It is clear how the amplitude falls to a low value near each critical frequency.

(iii) *Critical Frequency Phenomena for the Abnormal Region E*

Records taken near the critical frequency of the abnormal region E are shown in Figures 10 and 11. Unlike the examples just discussed, there is no group retardation observable on the (P',f) curves near the critical frequency. Also there is a range of frequencies for which simultaneous reflections are obtained from the abnormal region E and region F. In Figure 10 this range of frequencies is small (of the order of 20 kc/s.) while in Figure 11 it is larger (of the order of

300 kc/s.). These characteristics of the abnormal region E are well known, and there have been two main theories to explain the phenomenon.

The first theory may be called the 'scattering cloud' theory (Best, Farmer and Ratcliffe 1938, Appleton, Naismith and Ingram 1939, Appleton and Naismith 1939, 1940). In this theory it is supposed that clouds of high ionization density are present in region E which are capable of producing a scattered reflection on frequencies above the normal critical frequency. The 'overlap' between the E and F reflections is explained on this theory by region F being seen through the 'holes' between the clouds. In order to explain the fact that the reflections from region F can on some occasions be completely suppressed, it is suggested that the clouds can become sufficiently intense for the spaces between to be more or less filled up. The frequency on which echoes from region F first appear is interpreted as that frequency which can first penetrate the least dense portions of the patchy layer, and the frequency on which the reflections from region E finally disappear is that frequency which is capable of penetrating the most densely ionized portions. Recently, the theory has been developed further by Booker (1950), who has suggested that the frequency on which the echoes from region E disappear may be related to the *size* of the ion clouds rather than to their maximum ionization density.

The alternative theory of the abnormal E phenomenon may be called the 'thin-layer' theory. This seeks to explain the fact that simultaneous reflections may be obtained from regions E and F by invoking the partial reflection and partial transmission effects which occur when the refractive index of a medium changes appreciably in a distance comparable with the wavelength. Kirby and Judson (1935) suggested that the required sharp gradient of refractive index might occur at the base of the normal region E. As this suggestion is unacceptable in its simple form, later workers have assumed that an extra layer of ionization exists near the maximum of the normal region E. If this layer is thin, the presence of the simultaneous reflections from regions E and F and the absence of group retardation at the critical frequency can be explained.

The variations of reflection coefficient and transmission coefficient as functions of frequency for such a 'thin' layer have been investigated theoretically by Hartree (1929), Booker (1938), Rydbeck (1942, 1944), Rawer (1939, 1940) and others. The problem requires a full wave treatment, and a particular shape of layer must be assumed. If absorption is neglected, and if a parabolic region is assumed, of semi-thickness $y_m > \lambda/2\pi$, the reflection coefficient ρ for frequencies near the critical frequency (f_c) is given by

$$\rho = [1 + \exp\{(4\pi^2/\mathbf{c})y_m(f - f_c)\}]^{-1/2}. \qquad \ldots\ldots(7)$$

The study of the amplitudes of the reflected and transmitted waves as functions of frequency by means of the rapid frequency sweep apparatus can give some information about the nature of the abnormal E reflections. Consider first the record shown in Figure 11. If one attempts to explain the large overlap between the reflections from regions E and F on the hypothesis of a thin layer, the thickness of the abnormal region E is found to be a few hundred metres. Some workers have considered such small thicknesses to be real. It is clear, however, that if the layer is really so thin, the perfect horizontal stratification, which is necessary for the thin layer theory to be applicable, will be very difficult to maintain. Figure 11 shows that the amplitude fluctuates in an irregular manner as the frequency is varied, and this must mean that the ionization is irregular in the horizontal plane. An inspection of all the records shows that when the 'overlap' is large pronounced irregularities are always present. Records of this type, then, favour the hypothesis of scattering clouds. In extreme cases, a very scattered type of reflection can occur over most of the frequency range, with little variation of 'reflection coefficient'. On the other hand, the absence of random fluctuations in Figure 10 shows that the irregularities of ionization are small, and the assumption of horizontal stratification is justified. If we interpret this record on the hypothesis of a thin layer, the total thickness of the region is found to be about 7 km. Thus, it appears that both theories may be needed to explain all the observed effects.

An examination of records taken on about 40 different occasions shows that they tend to divide themselves into the two types illustrated by Figures 10 and 11, suggesting that two distinct types of abnormality can be present in region E.

It is of interest to see whether the two types of abnormality could be detected by other methods. We might expect that the 'irregular' type would be characterized by a reflected wave which had components coming from angles considerably removed from the vertical, while the 'thin layer' type would not show such a wide spread. Now we have already seen that the angular spread of the ionospherically reflected wave can be measured by experiments using spaced receivers. Briggs and Phillips (1950) have given some results obtained in this way for the abnormal region E. It was found that a histogram of the angular spread θ_0 had two peaks, one at 6° and the other at 20°. These results were for a frequency of 2·4 Mc/s. From equation (5) we can calculate the number of maxima to be expected in a given frequency range for these two values of θ_0. Taking $h = 120$ km., and $\Delta f = 500$ kc/s. we find $\Delta n = 2{\cdot}2$ and 51 for the two cases. The value of 6° could therefore be taken as representative of the type of record shown in Figure 10, and the value of 20° would correspond to a record like that shown in Figure 11. Strictly these calculations would only apply to a frequency band centred on 2·4 Mc/s., but the order of magnitude of the effects will be similar for any frequency band not too far removed from this frequency. Thus, although the two peaks in the histogram referred to above were not well marked, these results do tend to confirm the hypothesis that two types of abnormality can occur, and there is reasonable agreement between the two experiments.

In view of these results, it is interesting to note that two different causes of abnormal E ionization have been proposed (Appleton and Naismith 1947; the theories have been summarized by Lovell 1946). The extra ionization causing the abnormal effects is believed to be associated partly with meteors and partly with magnetic activity. Perhaps the 'rough' type, illustrated by Figure 11, may be due to 'scattering clouds' of meteoric origin, and the 'smooth' type, illustrated by Figure 10, may be the 'auroral' type of abnormal E associated with magnetic activity. Attempts to correlate the occurrence of the two types with known cases of meteoric or magnetic activity have been unsuccessful. Indeed, it must be admitted that very scattered reflections of the abnormal E type are often observed during intense magnetic storms. However, even the normal ionospheric regions can give scattered reflections during such intense disturbances, so that this observation is not necessarily typical of the abnormal region E on such occasions.

It is of some interest to find the thickness of the abnormal region E on those occasions when the 'thin layer' theory is applicable. The amplitude of the echo from region E gives the reflection coefficient of the layer, and the amplitude of the echo from region F gives the *square* of the transmission coefficient. As the height of region F is approximately twice that of region E, the echo which has suffered two reflections from region E sometimes falls within the strobe which has been adjusted to select the echo from region F, and care must be taken to avoid incorrect results due to this cause. An analysis of the records shows that there is usually considerable absorption near the critical frequency, and so equation (7), which neglects the effects of absorption, is not strictly applicable. Rydbeck and Rawer have considered theoretically the case of a thin layer *with* absorption. A few of the records are so free from random fluctuations that an accurate comparison with the theoretical results can be made, and it is then possible to deduce both the thickness of the layer and the value of the collision frequency (Briggs 1951). Normally there are small fluctuations even in the records which we have classed as 'smooth', and one is not justified in attempting to deduce anything more than the order of magnitude of the thickness of the layer. For this purpose, equation (7) may be used, and the value of y_m determined from the observed variation of reflection coefficient with frequency. Figure 12 shows a histogram of the values of the total thickness of the layer $2y_m$ obtained in this way. The smallest value observed was 1 km.

§ 5. EXPERIMENTAL RESULTS: PANORAMIC RECORDS

Panoramic records produced by the use of the Type D display and a continuously moving film, show the fluctuations of the amplitude of a selected echo over a fixed frequency band as time progresses. In Figures 13 to 24 short sections of these records are shown. These have been chosen to illustrate the various types of record which are obtained for different ionospheric conditions.

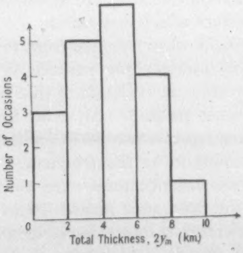


Figure 12. Histogram of observed thicknesses of the abnormal region E for occasions when the hypothesis of a 'thin layer' is applicable.

The sections of the records which are reproduced each cover a time of approximately 12 minutes, and the frequency range is shown on each record. The short gaps are timing marks occurring at ten-minute intervals. There is a cut-off effect at low signal levels due to the characteristics of the cathode-ray tube and the film; thus a completely black area on the records does not necessarily mean that the amplitude of the echo was zero.

(i) *Records showing Random Fluctuations of Amplitude*

If the ionosphere behaved like a specular reflector, the panoramic records would show simply the same amplitude on all frequencies and at all times. Again, if the ionosphere behaved like an ideal 'random reflecting screen', free from any critical frequency effects, the records would show an irregular pattern of constant statistical properties. Such records are often obtained on occasions when no critical frequency effects are present. The degree of roughness may be very different on different occasions. Figure 13 shows some short samples of records of this type, arranged in order of increasing roughness (*a*) to (*f*) (Plate II).

The amplitude R of the reflected wave is a function $R(f, t)$ of the two variables frequency f and time t. A panoramic record is a representation of the function $R(f, t)$ on a two-dimensional film. The structure of the record in the frequency direction depends upon the number of maxima of R which occur in a given frequency band. This was discussed in § 3, and it was shown that for reflection at a single horizontal layer, the number of maxima was given by equation (5). The structure of the panoramic record in the *time* direction depends upon the number of maxima of R per unit time, in other words, upon the *fading speed*. This point was also discussed in § 3 and the expression (6) was obtained for the number of maxima occurring per unit time. From a comparison of equations (5) and (6), we see that an increase of θ_0 will cause an increase of the number of maxima in both the frequency and time directions. An increase of the horizontal velocity V will cause an increase in the number of maxima in the time direction, but will leave the structure in the frequency direction unaffected. An inspection of the records shows that the main changes which occur are due to changes of θ_0, which affect the structure in both directions. The magnitude of V does change from time to time, but the changes are less important than the changes in θ_0. In other words, it is changes of the 'roughness' of the reflecting region which produce the main differences between the records obtained on different occasions.

Even when no critical frequency phenomena are involved, the records are not always of a completely 'random' type. For example, Figures 14 and 15 show two different types of structure which may occur.

Figure 14 is a record of the wave reflected from region E and the lines which cross the record probably indicate the pesence of clouds beneath the main layer, which is itself fairly smooth. Clouds of this type have been detected by Findlay (1947) by a different method. An echo from a cloud and from the layer itself would produce a typical 'two path' condition, and we should expect periodic variations of amplitude as the frequency is changed. If the path difference between the two contributions were constant, there would be no variation with time, and the panoramic record would show lines parallel to the time axis. If the path difference is increasing or decreasing due, for example, to a horizontal movement of the cloud, it is easy to see that lines will cross the record diagonally, as is the case in Figure 14.

The record in Figure 15 is for region F. Records of this type, showing a curious structure, are often obtained for this region, and no explanation has been found, though the effects may be related in some way to the proximity to the F1 critical frequency. The possibility of ion clouds in region E affecting the reflections from region F must also be considered in an attempt to interpret records for this region.

(ii) *Critical Frequency Phenomena*

Owing to the limited frequency sweep which can be used with the present apparatus, the panoramic method has not been found to be very useful for studying the fluctuations of the critical frequency of region F. It can, however, give useful results for region E. The most interesting effects relate not to the main critical frequency of the region but to the minor critical frequencies sometimes present below the main one. These are associated either with small maxima in the curve of ionization against height or, more generally, with points at which the ionization is constant over a small range of heights. Such irregularities in the curve of ionization against height have been called 'ledges', and it is convenient to extend this term to include the characteristic effects produced on the records.

The ledges may be well developed, as in the example of Figure 9, or much less pronounced. In all cases there is absorption near each of the critical frequencies, so that on the panoramic records these frequencies appear as dark lines of weak signal running along the records. This is a very sensitive method of detecting the presence of ledges. In some cases very little change of P' can be detected on a (P',f) record, but the presence of the ledge shows up clearly on a panoramic record. Whale (1951) has investigated the same effects by means of a careful study of (P',f) curves. The present work is complementary to his work, the critical frequencies being detected by the associated absorption rather than by the changes of P'.

Figure 16 shows the presence of a very well-marked critical frequency of the type now being considered. The line of weak signal is clearly visible, and is of the order of 100 kc/s. in width. Figure 17 shows two rather less pronounced ledges appearing on the same record. Figure 18 is an example of a minor form of the phenomenon in which the absorption is not very great at the critical frequency.

Effects of this general type are found to be present for about 15% of the total time for which echoes are obtained from region E in the frequency band 2·2–2·6 Mc/s.

The critical frequencies associated with the ledges are always found to be *decreasing* in frequency whatever the time of day at which they are observed. This can be seen in the three examples reproduced. The rate of decrease of critical frequency is usually about 10 kc/s. per minute. The same effect has been observed by Whale, show finds that the discontinuities in the (P',f) curve always start at the critical frequency of region E and move down the curve towards lower frequencies. This could be explained by a downward movement of a thin layer of ionization. It is interesting to note, however, that Munro (1949, 1950) has observed a rather similar effect in region F, and Martyn (1950) has suggested that the apparent movement of a perturbation down the (P',f) curve is not real, but is due to a disturbance travelling horizontally with a tilted wavefront. It may be that the effects in region E are of a similar nature to those discussed by Martyn and Munro for region F.

Another interesting feature which sometimes appears on the panoramic records is that the degree of irregularity of the record is greater on frequencies above the critical frequency of a ledge than on frequencies below it. This effect is clearly visible in Figure 18. It is found to occur on about half the occasions on which ledges are observed. Since the level of reflection in the ionosphere rises as the frequency increases, this result must mean that the ionization above the height of the ledge is more irregular than below this height. Figure 19 shows another form of this phenomenon of less frequent occurrence, in which there is a change of structure without any absorption at the transition frequency. In this case we must have simply an increase of the degree of irregularity above a certain height, not associated in any way with a critical frequency. These results show that irregularities in the ionosphere may be sharply localized in height.

We shall consider next some records which show fluctuations in the critical frequency of the abnormal region E. Figure 20 is a typical example in which the critical frequency is fairly well defined. At one point the selecting strobe was changed so that the echo from region F was recorded for a short time. It will be seen that in this case there was no appreciable overlap between the echoes from regions E and F. The portion of the record which shows the amplitude of the echo from region F has an interesting structure, probably produced by diffraction effects in region E.

Another record is shown in Figure 21. At the start of this record, the critical frequency is very well defined, and the record shows few irregularities. Then an additional echo appears on a frequency above the critical frequency and gradually spreads in frequency range until it joins up with the main echo. This phenomenon is repeated a few minutes later. Finally, the critical frequency is found to have increased, and the irregularity of the record is also greater than before.

The two records just discussed have shown the upper frequency limit, or critical frequency, of the abnormal region E. The low frequency limit for echoes from this region occurs at the region E 'cusp' of the (P', f) curve; frequencies below the cusp are reflected from the normal region E, and above the cusp from the abnormal region E. Near the cusp there is group retardation and absorption associated with the critical frequency of the normal region E. Figure 22 shows a panoramic record embracing the cusp frequency, which occurs about the centre of the frequency sweep. The low-frequency limit of the reflections from abnormal E can be seen near the right-hand side of the record; this frequency is very constant and well defined. The weaker reflections having a stratified appearance are from the normal region E. Occasionally one of these strata may begin to develop and move away from the cusp frequency towards lower frequencies, thus forming one of the ledges discussed above. Unfortunately, it is not possible to show this effect on a short length of record.

The variety of effects associated with the abnormal region E is very great, and it is only possible here to give these few examples as an illustration of the value of the panoramic method of recording.

(iii) *Records showing Interference between the Two Magneto-Ionic Components*

All the records so far discussed represent the behaviour of the ordinary ray selected by the polarized aerial. If it is desired to study the interference effects produced by admitting both the magneto-ionic components, a simple dipole aerial may be used. Figures 23 and 24 show two records obtained under these conditions. Occasions were selected when the two components were present with approximately equal strengths. Figure 23 is a record of reflections from region F for a frequency range well removed from the critical frequency. Figure 24 is a record of reflections from the abnormal region E, and the critical frequency of the ordinary ray occurs near the centre of the record.

The main phenomenon of interest is the periodic variation of amplitude as the frequency is varied. The periodicity in the frequency direction provides an accurate means of measuring the difference of the group paths of the two magneto-ionic components. To demonstrate that the periodic variation of amplitude has, in fact, been caused by the rotation of the plane-polarized wave formed by the addition of the two circularly polarized components, the receiver may be switched for a short time to the second aerial of the crossed pair. This was done once in the record shown in Figure 23 and twice in Figure 24. It will be observed that the positions of the maxima and minima are exactly interchanged, as expected.

§6. THE MEASUREMENT OF COLLISION FREQUENCIES OF ELECTRONS AT VARIOUS HEIGHTS

From records of the type illustrated by Figures 8 and 9, the collision frequency of electrons can be found by the use of the well-known formula (Appleton 1935):

$$\Delta(\ln \rho) = -(\nu/2\mathbf{c})(\Delta P' - \Delta P) \qquad \ldots\ldots(8)$$

which relates the change of reflection coefficient ρ to the changes of the group path P' and the phase path P. Changes of P are usually small compared with the changes of P' near the critical frequency of a region. Consequently, a graph of $\Delta(\ln \rho)$ against $\Delta P'$ gives a straight line from the slope of which ν can be determined. The height to which the value of ν refers depends on the variation of ν with height, but in practice it can usually be taken as the height at which the ionization density has its maximum value.

It cannot be assumed that a single (R,f) record will give the variation of reflection coefficient with frequency directly, because of the presence of the random fluctuations previously discussed. In order to obtain a (ρ,f) curve —reflection coefficient as a function of frequency—it is necessary to average a number of successive (R,f) records so that the random fluctuations are eliminated. In some cases the ionization is sufficiently regular for this procedure to be unnecessray, but the only method of demonstrating that such conditions hold is to record a number of (R,f) records and show that they do not change with time. The examples shown in Figures 8 and 9 are, in fact, fairly free from fluctuations of the random type, so that the single (R,f) records may be interpreted directly as (ρ,f) records.

Farmer and Ratcliffe (1935 a, b) have determined values of ν for region F by the use of equation (8). With the more rapid recording which is possible with the apparatus described in this paper, it is now possible to extend the method to region E, where the complex penetration effects are on too fine a scale and too rapidly changing to permit the use of the manual methods used in the earlier experiments. Records of the type shown in Figure 9 enable the value of ν to be found at various heights between 100 km. and 150 km. A full analysis of the records is now in progress. Preliminary results show that the value of ν at 130 km. is about $2 \times 10^4\,\mathrm{sec}^{-1}$.

ACKNOWLEDGMENTS

This work has been carried out as part of a programme of radio research at the Cavendish Laboratory supported by a grant from the Department of Scientific and Industrial Research. The author is also indebted to the same Department for the award of a maintenance grant.

REFERENCES

Appleton, E. V., 1935, *Nature, Lond.*, **135**, 618.

Appleton, E. V., and Naismith, R., 1939, *Nature, Lond.*, **143**, 243; 1940, *Proc. Phys. Soc.* **52**, 402; 1947, *Ibid.*, **59**, 461.

Appleton, E. V., Naismith, R., and Ingram, L. J., 1939, *Proc. Phys. Soc.*, **51**, 81.

Best, J. E., Farmer, F. T., and Ratcliffe, J. A., 1938, *Proc. Roy. Soc.* A, **164**, 96.

Booker, H. G., 1938 (quoted by Appleton, Naismith, and Ingram, 1939); 1950, *Conference on Ionospheric Physics*, Penn. State Coll., July 1950.

Booker, H. G., Ratcliffe, J. A., and Shinn, D. H., 1950, *Phil. Trans. Roy. Soc.* A, **242**, 579.

Breit, G., and Tuve, M., 1926, *Phys. Rev.*, **28**, 554.

Briggs, B. H., 1951, *J. Terr. Atmos. Phys.*, in the press.

Briggs, B. H., and Phillips, G. J., 1950, *Proc. Phys. Soc.* B, **63**, 907.

Briggs, B. H., Phillips, G. J., and Shinn, D. H., 1950, *Proc. Phys. Soc.* B, **63**, 106.

Farmer, F. T., and Ratcliffe, J. A., 1935 a, *Nature, Lond.*, **135**, 485; 1935 b, *Proc. Roy. Soc.* A, **151**, 370.

Findlay, J. W., 1947, *Nature, Lond.*, **159**, 58.

Hartree, D. R., 1929, *Proc. Camb. Phil. Soc.*, **25**, 97.

Kirby, S. S., and Judson, E. B., 1935, *Proc. Inst. Radio Engrs., N.Y.* **23**, 733.

Lovell, A. C. B., 1946, *Rep. Prog. Phys.*, **11**, 415 (London: Physical Society).

Martyn, D. F., 1950, *Proc. Roy. Soc.* A, **201**, 216.

Moxon, L. A., Croney, J., Johnston, W. G., and Laws, C. A., 1946, *J. Instn. Elect. Engrs.*, **93**, Pt. III A, 1143.

Munro, G. H., 1949, *Nature, Lond.*, **163**, 812; 1950, *Proc. Roy. Soc.* A, **202**, 208.

Phillips, G. J., 1951, *Proc. Instn. Elect. Engrs.*, **98**, Pt. III, in the press.

Rawer, K., 1939 a, *Hochfrequenztech. u. Elektroakust.*, **53**, 150; 1939 b, *Ann. Phys., Lpz.*, **35**, 385; 1940, *Naturwissenschaften*, **28**, 577.

Rydbeck, O. E. H., 1942, *J. Appl. Phys.*, **13**, 577; 1944, *Trans. Chalmers Univ. Tech., Gothenburg, Nr.* 34.

Whale, H. A., 1951, *J. Terr. Atmos. Phys.*, in the press.

LETTERS TO THE EDITOR

Infra-Red Photoconductivity of certain Valence Intermetallic Compounds

The high sensitivity, but limited spectral range, of photoconductive materials such as Tl_2S, PbS and PbTe has led to a search for similar materials, sensitive to radiation of longer wavelength. Moss (1950) has surveyed existing results and attempted a theoretical approach to the problem. In the present investigation further materials of this class (called ' valence intermetallic compounds ' by Hume-Rothery (1944)), have been studied.

The materials were prepared by heating together appropriate quantities of the constituent elements in silica crucibles with well-fitting lids. The fused materials obtained were powdered in a mortar. Layers were prepared from these powders by evaporation *in vacuo* of the order of 10^{-5} mm. Hg in Dewar type cell blanks. Sensitivity spectra were obtained using a Leiss monochromator the output of which was calibrated with a Schwartz type thermocouple whose response was assumed to be independent of wavelength.

With one exception the results given are for layers which had received no special sensitization treatment after being formed. The exception, Tl_2Te, is analogous to Tl_2S, which has been studied very thoroughly by von Hippel *et al.* (1946). Layers of Tl_2Te as evaporated have a dark colour, low electrical resistance and no sensitivity. On heating in air pressures of the order 10^{-1} mm. Hg the layers turn yellow and have very high resistances. With care it is possible to obtain layers in an intermediate state which are photosensitive if cooled with liquid oxygen. Some attempts were made to improve the responses of layers of other materials by treatment of this sort, but without success.

Tellurium is itself a photoconductor (Moss 1949), so that free tellurium may affect the spectra of tellurides. If the telluride evaporates at a different temperature from tellurium, the latter can be distilled off during the evaporation process, but there is still the possibility that the telluride dissociates during evaporation. Usually the spectra obtained showed no similarities to the spectrum characteristic of tellurium, leaving little doubt that any effect of free tellurium was unimportant, but for a few materials the spectra varied from layer to layer, suggesting that the amount of free tellurium present varied from layer to layer. Zinc, cadmium and tungsten tellurides behaved in this way.

The materials studied are listed together with the long-wavelength threshold, which is defined as that wavelength at which the response has fallen to 50% of its peak or plateau value. All the results are for layers cooled with liquid oxygen except Zn_3As_2, which was measured at room temperature. This material was sensitive when cooled with liquid oxygen, but its resistance was extremely high. It has a subsidiary sensitivity band superimposed on the main threshold, figures for both thresholds being quoted.

(1)	(2)	(3)	(4)
Cu_2Te	—	1·2	2·2
Ag_2Te	—	~1·3	3·0
ZnTe	0·8	1·4	3·5
HgTe	2·6	3·1	3·9
Tl_2Te	1·2	1·6	2·6
Sb_2Te_3	—	1·6	2·6
Mo_2Te	0·9	1·6	2·6
W_2Te	1·0	1·35	—
U_2Te	1·0	1·35	2·5
Zn_3As_2 (room temperature)	—	1·1	
	1·25	1·4	1·9
SnS	—	0·9	1·1
Sb_2Se_3	—	0·95	1·2

(1) Material. (2) Peak wavelength (μ). (3) Long wavelength threshold (μ). (4) Longest wavelength at which signals were detected (μ).

Notes.

(*a*) The figures given for ZnTe and WTe_2 are for the layers of these materials which appeared to be least affected by excess or free tellurium.

(*b*) The spectra of Mo, W and U tellurides are fairly narrow bands, apparently similar to the spectrum of artificially prepared MoS_2 as described by Hughes and Dubridge (1932). According to these writers the 2μ threshold quoted by Moss (1950) occurs only in natural crystals of MoS_2.

The sample of UTe_2 used was very kindly provided by Dr. R. H. Myers of the Atomic Energy Research Establishment, Harwell. The work described was carried out at the Telecommunications Research Establishment, and acknowledgment is made to the Chief Scientist, Ministry of Supply, and to the Controller of H.M. Stationery Office, for permission to publish this note.

Telecommunications Research Establishment,
Ministry of Supply, Malvern.
9th January 1951.

J. G. N. BRAITHWAITE.

VON HIPPEL, A., *et al.*, 1946, *J. Chem. Phys.*, **14**, 355.
HUGHES, A. L., and DUBRIDGE, L. A., 1932, *Photoelectric Phenomena* (New York: McGraw-Hill Book Co.).
HUME-ROTHERY, W., 1944, *Structure of Metals and Alloys*, Part V (London: Institute of Metals).
MOSS, T. S., 1949, *Proc. Phys. Soc.* A, **62**, 264; 1950, *Proc. Phys. Soc.* B, **63**, 167.

Heavy-Current Arcs in a Transverse Magnetic Field

Observations have been made recently on some properties of heavy-current arc discharges in hydrogen at low pressure in a transverse magnetic field. The discharges were pulsed, being of duration 2·5 milliseconds, and occurred at a known instant in a pulsed magnetic field of 10 milliseconds duration and sinusoidal pulse shape (Champion 1950).

By means of an oscillograph, measurements were made of the arc current and potential fall with two types of electrode configuration in magnetic fields of various strengths.

In the first case, the arc was passed between two tungsten rod electrodes, with their ends 2·5 centimetres apart, on a diameter of a cylindrical glass discharge tube of 4 centimetres diameter. At a pressure of hydrogen of the order of 5 mm. Hg, and with no magnetic field, an 80 ampere arc filled the cross section of the tube. A magnetic field constrained the discharge in a path around one side of the tube. The arc volume was reduced, and showed an increased intensity of light emission in the magnetic field.

The voltage–current characteristic of the arc showed a slight negative gradient over the range 25–80 amperes, with no magnetic field, becoming increasingly negative with increasing magnetic field. A plot of the ratio of arc potential drop to current (termed the 'effective impedance') against the magnetic field is of the form shown in Figure 1 for a pressure of 5 mm. Hg. The curves represent measurements taken at the mid-point in time of the pulse; equilibrium was then established.

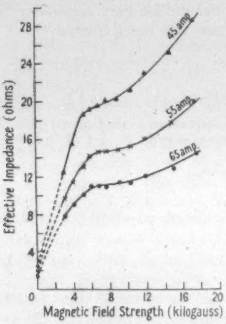


Figure 1.

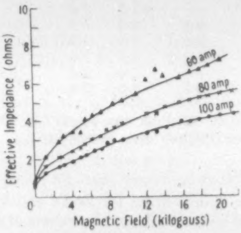


Figure 2.

A study of the curves shows: (i) the effective impedance increased with increasing pressure with no magnetic field, but decreased with increasing pressure in a magnetic field, (ii) at 5 mm. Hg pressure, the curves showed a region in which the impedance of the arc, and hence, power input at constant voltage, remained roughly constant with increasing magnetic field. This feature was less pronounced at lower arc currents and pressures. At 0·4 mm. Hg pressure it was absent for all currents used.

An increase in the effective impedance of the arc by the magnetic field is to be expected from the reduced mobility arising from the cycloidal motion of charged particles in crossed electric and magnetic fields. Theoretically (Townsend 1912) it is deduced that a field of 1,000 gauss reduces the diffusion coefficient, across the magnetic field, of electrons of temperature 20,000 degrees at 2 mm. Hg pressure to 1/260 of its value with no field. However, Langmuir and Mott-Smith (1924) point out that the mobility of an electron near the wall of a tube in the presence of a magnetic field is reduced but slightly compared with the mobility without a field. In the present case, the 'step' in the impedance curves was probably due to a pressure-dependent effect on ionizing processes near the wall of the tube.

In the second set of experiments, the discharge passed between a 'Nichrome' cylinder, 5·1 cm. long and 3·8 cm. in diameter, and a coaxial tungsten rod cathode of 2 mm. diameter. The magnetic field was paraxial with the electrode system. The arc, which was localized with no magnetic field, appeared visually to fill the whole electrode volume in the magnetic field. This may have been due to rotation of the arc about the axis.

Figure 2 shows values of the effective impedance of the arc plotted against the magnetic field strength, for hydrogen at a pressure of 5 mm. Hg. There was no evidence of a horizontal 'step' in these curves, and the effect of the walls of the containing vessel was considered negligible.

The arc voltage–current characteristic had a slightly positive gradient in the absence of a magnetic field, but it became more negative with increasing magnetic field.

In both experiments, the power dissipated in the discharge, to maintain a given current, increased with magnetic field. In the former case, the maximum instantaneous power dissipation was about 60 kw., and in the latter about 25 kw., for a current of 65 amperes at 5 mm. Hg pressure. Some power was lost to the wall of the tube in the former case, but in the latter case, it may be assumed that almost all the power was dissipated in the gas, and at the electrodes.

The writer wishes to thank Professor J. Sayers and Mr. K. S. W. Champion for their encouragement in this work, which was made possible by a grant from the Department of Scientific and Industrial Research.

Electron Physics Department,
The University, Birmingham.
8th January 1951.

N. L. ALLEN.

CHAMPION, K. S. W., 1950, *Proc. Phys. Soc.* B, **63**, 795.
LANGMUIR, I., and MOTT-SMITH, H., 1924, *Gen. Elect. Rev.*, **27**, 810.
TOWNSEND, J. S., 1912, *Proc. Roy. Soc.* A, **86**, 571.

Discussion

on paper by H. K. HENISCH and J. EWELS entitled "A Study of Electrical Forming Phenomena at Selenium Contacts" (*Proc. Phys. Soc.* B, 1950, **63,** 861); *read at Science meeting of the Society on 24th November* 1950.

Dr. P. T. LANDSBERG (*Department of Natural Philosophy, University of Aberdeen.*) I regret that I cannot offer any more than a rather general remark about this paper. It concerns the quantitative theory of these time-dependent processes and, although my results are quite incomplete, an outline of the ideas involved may possibly be of interest. A theory of the time-dependent current–voltage curves may be envisaged as involving three steps. (i) The characteristic must be obtained for an arbitrary distribution of space charge in the barrier. Using known methods (Landsberg 1949, 1951), and assuming reverse or small forward potentials $U/\boldsymbol{e}$, the current density j can in fact be determined explicitly in the form

$$j=j_0fe^y[1-\exp(U/\boldsymbol{k}T)],$$

where y is a function of the space charge distribution (and hence of the applied potential), the temperature T, and the dielectric constant of the semiconductor; j_0 depends on the physical constants of the junction, but is independent of the applied voltage; $f=1$ for the diode and $y^{3/2}$ for the diffusion theory. (ii) The likely time-dependence of the space charge distribution must be studied with the aid of suitable models. (iii) A quantitative theory of the time-dependent characteristics should follow by combining (i) with (ii).

Mr. T. R. SCOTT (*Standard Telecommunications Laboratories, Ltd., Enfield, Middlesex*). While the authors have made a valuable contribution towards the study of creep phenomena and have proposed mechanisms, which, although only qualitative, appear to agree with many observed facts, I feel that it is somewhat early to formulate theories. Many other observed phenomena do not appear to be accounted for easily by their suggestions.

Our work has been mainly related to creep under A.C. conditions and while some phenomena observed are in line with the authors' observations, others are not. For example, under 50 c/s. conditions positive creep in a shelf-aged rectifier persists only for a few milliseconds after voltage application. Figure 12 suggests similar behaviour but the time scale is quite different and some explanation of the difference is required. Similarly under

A.C. conditions positive creep is profoundly affected by the presence of even very small forward currents. This effect must be brought into any complete theory.

Temperature effects are not entirely cleared up by the authors' experiments from −180 to +70° C. Work at present in progress at temperatures in the 100 to 140° C. region suggests that this may be a fruitful field of study. Also, while storage at elevated temperatures may improve the reverse characteristic, its effect on the forward characteristic may be quite different. Even the former needs qualification since the authors' time scale is such as to leave open to doubt the effects observed in the first few milliseconds. The distinction between the effects of self-heating and of ambient temperature would also appear to require some elucidation.

Finally, it is modern practice to assist the barrier layer by adding a thin film of insulant under the counter-electrode. This lacquer, in the authors' temperature range at least, almost completely eliminates positive creep. Its presence, however, appears to complicate the theory of the barrier layer.

AUTHORS' REPLY. We agree with Dr. P. T. Landsberg on the formal steps which should eventually lead to a quantitative theory of forming and current creep. The number of unknown parameters is, however, still very large and a good deal of further experimental work is necessary before such a theory could be established on a reliable basis.

The tentative nature of the present model is freely acknowledged. It accounts for the principal features of the results known to us from the literature and our own experiments, and, as new evidence becomes available, the model may have to be modified.

Mr. T. R. Scott's comments lead us to conclude that an investigation of creep under D.C. conditions during the first few milliseconds would be of interest for comparison with the corresponding A.C. measurements. Such a comparison would, of course, have to be carried out *for equal power dissipations*. Since even the temporary presence of forward currents materially affects the power dissipation, their influence on positive creep is not unexpected.

Further work on temperature effects over a wide temperature range would certainly be valuable. The exact mechanism of artificial barrier layers on deficit semiconductors is not yet understood. The appearance of positive creep is expected to depend on the height of the contact barrier, and no predictions are possible at present as to how this height is affected by the presence of the insulating varnish.

LANDSBERG, P. T., 1949, *Nature, Lond.*, **164**, 967; 1951, *Proc. Phys. Soc.* B, **64**, 82.

REVIEWS OF BOOKS

Radiative Transfer, by S. CHANDRASEKHAR. Pp. xiv + 393. 1st Edition. (Oxford: Clarendon Press 1950.) 35*s*.

In this book Professor S. Chandrasekhar, F.R.S., who holds the chair of theoretical astrophysics in the University of Chicago, makes an outstanding addition to the systematic treatises on astrophysical problems. Though the book is intended primarily for astronomers it also has great meteorological interest, because the methods of calculation of radiative transfer which it describes are applicable to the very difficult problem of multiple scattering of light in the earth's atmosphere; moreover, essentially similar problems occur in connection with the diffusion of neutrons, hence the book should also attract the attention of theoretical physicists. The main topic, however, is the radiative equilibrium of stellar atmospheres, treated with great breadth and also in much mathematical detail.

The subject goes back at least to Rayleigh's discussion (1871) on the blue light of the sky and its polarization; as far as the earth's atmosphere is concerned, further progress was made by L. V. King, and also by S. Chapman and A. Hammad, but they recoiled from the difficulties of a complete formulation of the problem and its solution; Chandrasekhar, in recent years, in a long series of papers in the *Astrophysical Journal*, has boldly faced and overcome these difficulties with great mathematical skill; one element in his success is the use and extension of a principle of invariance introduced by the Russian astrophysicist V. A. Ambarzumian (1943). Chandrasekhar, in one of a series of Bibliographical Notes which conclude each chapter (and form a valuable and distinctive feature of the book), points out that some of the basic ideas involved resemble those introduced by Stokes and Rayleigh in the discussion of the reflection and transmission of light by thin plates. Among other authors whose names occur in these Notes are Schuster, Schwarzschild, Sampson, Eddington, Milne, Hopf, van de Hulst and Crum.

The book has a good subject index and index of definitions, but lacks an authors' index.

Radiative transfer is a subject of considerable intrinsic mathematical complexity, but this brilliant and powerful book will smooth the path of knowledge for many readers and thereby give a new impulse to further development of the subject. S. C.

Photoelectricity and its Application, by V. K. Zworykin and E. G. Ramberg. Pp. vi+494. (New York: John Wiley and Sons; London: Chapman and Hall, 1949.) 60*s*.

The names of the authors are sufficient recommendation. The senior author is the most eminent living inventor in the applications of photoelectricity, and the junior author has already made a good name for himself, particularly by his work in electron optics.

The main emphasis in this book is on applications. Less than about one-quarter is given to the physics of photoemission, photoconductivity and barrier-layer effects, the rest is devoted to a description of the almost innumerable varieties of photoelectric devices and their utilization. The treatment is not only complete but a little too encyclopaedic. It leaves one wondering what sort of reader the authors have contemplated: probably one whose library consists of this single volume. Why else should they think it necessary to include seven pages on vacuum pumps and on the production and measurement of high vacua and ten pages on the elements of electron optics ? Such 'encyclopaedia articles' will be insufficient for the physicist who wants to make photocells, and they are not indispensable for the engineer who only wants to use them. But it must be admitted that it is impossible to write a book on a subject like this which should be equally satisfactory to the physicist, the engineer and the customer. Perhaps it is the last class of reader who gets more than his fair share, as the 'applications' part of the book is an admirably complete list of almost all devices on the American market which contain photoelectric elements of some sort.

The comprehensiveness of the book may be judged from its chapter headings: Introduction, General theory, Photosensitive surfaces, Materials and apparatus for making photocells, General methods for preparing phototubes, Vacuum phototubes, Gas-filled phototubes, Multiplier tubes, Image tubes, Photoconductive cells, Photovoltaic cells, Photocell circuits, Measurements of small photocurrents, Photoelectric measuring devices, Phototubes in sound reproduction, Phototubes in picture transmission, Photosensitive camera tubes in television, Light beam signalling, Miscellaneous applications, Photocells in the future, with a useful Appendix and excellent subject and author indexes.

Like its predecessor, *Photocells and their Applications*, by Zworykin and Wilson (second edition, 1934) this book will be indispensable for works and laboratory libraries in the electrical industry, for University libraries, and for every research or development engineer working in this field. D. GABOR.

The Properties of Metallic Materials at Low Temperatures, by P. Litherland Teed. Pp. 220. (London: Chapman and Hall, 1950.) 21*s*.

This book, the first of a series to be published under the auspices of the Royal Aeronautical Society, is a critical compilation of the available data on the subject indicated by its title. It is intended primarily for engineers and designers (all data have been assiduously converted to F.P.S. units), and tables of results and references occupy more than half of the text. The author is at pains to point out, however, how much more work needs to be done before a comprehensive picture of the subject can be presented, and how often an otherwise excellent piece of research has been spoiled by a failure to include an adequate description of the material on which the measurements were made.

The physicist interested in such matters has only to open the book at random to be presented with any number of problems awaiting solution—if he should ever find himself in the unlikely situation of requiring such a stimulus.

The book is very clearly printed and provided with a copious index: it should prove a valuable source of reference. N. THOMPSON.

An Introduction to Luminescence of Solids, by H. W. LEVERENZ. Pp. xv+569. 1st Edition. (New York: John Wiley and Sons; London: Chapman and Hall, 1950.) 96*s*.

This book is published in the 'Structure of Matter' series, and the author's intention is to provide "an orderly introduction to a growing field" for scientists of other interests. He also attempts to systematize the terminology of the subject and to provide a general coordination of luminescence phenomena, using specific materials as examples. The book begins with two introductory chapters, the first of which covers the elementary concepts of modern physics but includes rather irrelevant topics, such as Dirac's theory of positrons and pair production. Chapter II is more useful as it provides an introduction to the crystalline state with some specific reference to the crystal structure of typical luminescent solids. Chapter III is concerned with syntheses of well-known types of phosphors and the system of notation for defining the final products. Chapter IV discusses phosphor constitution, crystal structure, impurity activators and the electronic energy states of phosphors ; it includes a very brief discussion of photoconductivity.

Chapter V consists of more than two hundred text pages and covers all the physical phenomena encountered in the study of luminescence in solids. The space given to each topic is rather inadequate and better treatment of such important aspects of the subject might have been possible by affording them the space occupied elsewhere in the text by more speculative discussions. It will be difficult for those 'not skilled in the art' to distinguish the processes of luminescence treated in this chapter with respect to their relative importance, as the author's style is not very lucid. The next chapter provides a summary of phosphor characteristics, or, as the author terms it, 'a capsulized description of phosphors', while the last chapter is concerned with the various applications of phosphors. Five appendices and a glossary of symbols follow. Only the first appendix, dealing with the preparation of pure zinc and cadmium sulphides and selenides, and the third, giving useful conversion factors for photometric units, appear to be of definite usefulness. Some six hundred text references and a supplementary list of references to very recent publications are given which, apart from one or two important omissions, provide a valuable guide to the literature of luminescence since the titles of individual papers are included. The book concludes with adequate formulae, author and subject indexes.

Taken as a whole, the book contains a large amount of information on phosphors, particularly those of commercial interest, but its arrangement does not appear to be the best that could have been chosen. The reader may well object to the many cumbersome phrases suggestive of the jargon of luminescence laboratories but not helpful to the average scientific worker. Such phrases as 'consciously added activator', 'difficultly eradicable complex absorptions' and 'energy *transmittal* in solids' may be included with the above. For the industrial scientist the book may be of considerable interest as it represents the contribution of a recognized worker in applications of luminescence, who has set very high standards of purity and control in the production, and faithful reproduction, of phosphors of high quality to meet the stringent demands of commercial users. G. F. J. GARLICK.

The Mathematical Theory of Communication, by C. E. SHANNON and W. WEAVER. Pp. 117. 1st Edition. (Urbana, Illinois: University of Illinois Press, 1949). $2.50.

This book appears at a particularly opportune moment. The demands made by the increasing traffic on existing communications channels and the growing activity and interest in applying new methods of modulation, in particular pulse-code modulation, stress the urgency of the need for a new fundamental approach to the study of communication. The authors have been modest in their statements concerning the interest in this subject ; a vast technique and theory has developed around the science of electrical communication which is of application to many branches of science. Of recent years there has been a steady increase in the awareness of this fact, due, principally to developments in digital computing machines and in certain fields of biology. Drs. Shannon and Weaver have written this book from the point of view of the electrical communication engineer but, in common with much of the best modern work in this field, have used the language of statistical physics, as is required by the fundamental character of the subject.

In the main section of the book Dr. Shannon has condensed the work of various people into a coherent theory and has pursued certain concepts to their logical conclusions. The essential aim, at the outset, is to obtain a satisfactory mathematical definition of 'information'; it is now generally agreed that the 'information' (not to be confused with 'meaning' or with the statistician's use of the word) conveyed by a symbol essentially involves *selection* of that symbol out of a whole group of symbols having certain probabilities of selection P_i, and that the 'information' may reasonably be defined as $H = -\Sigma P_i \log P_i$ when the selections are independent. That 'information' expressed thus corresponds to a definition of the entropy of a system having states of probability $P_1 \ldots P_i \ldots P_n$, was perhaps first recognized by Szilard (*Z. Phys.*, **35**, 840) in 1929, in connection with the problem of the Maxwell demon. Physically this is reasonable, since information can exist only inasmuch as there is some 'uncertainty' in a message.

Dr. Shannon develops this thesis. In the first chapter he gives a description of a written message or symbol-sequence as a stochastic process, which is particularly clear. Taking the known frequencies of letters, digrams, trigrams and words as they occur in the English language and their known transitional probabilities, he illustrates the building up of letter and word sequences, obtaining extra clarity by the use of topological graphs, and thus develops the idea of a stochastic source of messages possessing a calculable 'entropy per symbol'. The idea of redundancy in a language is considered, zero redundancy corresponding to maximum entropy (since then there is nothing known about the message, *a priori*). The statistical structure of a language may be altered by encoding of the message source, and if an ideal coding can be found which maximizes the entropy this may be considered to 'match' the source to the channel (having a given capacity in symbols per second) and thus give the maximum rate of transmission of information, at least if the channel is noiseless.

Chapter 2 extends the theory to channels with perturbing noise, and develops the important theorem that in a given case there exists a code which, by the introduction of a certain redundancy, is optimal in combating noise, though it is admitted that this may be difficult to apply practically and no ideal coding method may in general be found. Chapter 3 proceeds to apply many of the concepts and theorems which refer to such channels (using discrete symbols) to channels with continuous waveforms, showing that no vital distinctions exist. Ensembles of such waveforms become analogous to a discrete symbol language, if they are 'sampled' at discrete intervals. The author's treatment of the representation of such sampled ensembles as a point distribution in multi-dimensional information-space, having a certain probability density, is very clearly presented.

The first criticism is that this book is over-condensed ; in spite of this the mathematical engineer, possessing the elements of probability theory, will glean much in the way of *general* theorems and concepts. Secondly, it might have been preferable to have put the section by Dr. Weaver at the beginning, rather than the end, as being essentially a broad introduction to the subject, straightforward in style, so that the comparative beginner might learn to swim, so to speak, by paddling first. As it is, he is taken straight to the deep end and thrown in.

COLIN CHERRY.

The Physics of High Pressures, by P. W. BRIDGMAN. Pp. vii + 445. (London: G. Bell and Sons, 1949.) 35*s*.

The new impression of Professor Bridgman's book with its supplement, consisting of a general survey of the recent high pressure field and a chronological survey of his own recent work, will again serve to draw attention to methods and techniques which are now being extensively used for investigation into the properties of matter at extreme pressures.

It will be recalled that in the original 1931 impression the author described the apparatus and methods he had employed for producing and measuring pressures up to about 20,000 kg/cm^2, the upper limit being set mainly by the mechanical properties of the steels then available for constructing the highly stressed members of the apparatus. In the intervening years advances in metallurgy have led to the development of a wider range of steels with enhanced mechanical properties and to the production of alloys much stronger than steel. Of these mention may be made of Carboloy, a tungsten carbide cemented with cobalt, which has a strength in straight compression of well over 60,000 kg/cm^2.

With the aid of these new materials and certain modifications in the design and construction of his pressure vessels, Professor Bridgman has now succeeded in carrying out an extensive series of measurements of physical changes up to 100,000 kg/cm^2, and of producing even higher pressures for short periods.

The accurate measurement of pressures in this new and extended range has an important bearing on the significance of the data obtained. Up to 30,000 kg/cm^2, pressure can be transmitted by a true fluid at room temperatures, and a secondary gauge, such as the manganin resistance gauge, can therefore be used provided a sufficient number of pressure-fixed points are available for checking the pressure coefficients of the operating element. One fixed point at 7,640 kg/cm^2 is given by the freezing pressure of mercury at 0° which can be determined to an accuracy within about 0·1 per cent with a primary free-piston gauge. A second suitable point is the I–II transition of bismuth at 30° which takes place at 25,420 kg/cm^2. Over this range the pressure–resistance curve for manganin has been determined by passing a second-degree curve through the two points. Beyond 30,000 kg/cm^2 the only method of measurement at present available is to determine the thrust on a piston subjected to the pressure, making corrections for frictional losses and distortion. The accuracy of the determination is estimated at within 2 or 3 per cent.

Over this range of pressure, the author has carried out volume measurements for a number of representative organic liquids and water at temperatures between 25°and 175°, and for a series of the more compressible elements. The melting curves of some twenty-one organic compounds and water have also been determined at temperatures up to 200° and at pressures up to 50,000 kg/cm^2.

The most striking feature in the pressure–volume–temperature relations of liquids is the approximate similarity of behaviour of all substances in the upper end of the pressure range. The changes of relative volumes between 5,000 and 50,000 kg/cm^2 all lie between 0·20 and 0·25 whether or not freezing is included. The volumes of both liquid and solid decrease along the melting line, the effect of increasing pressure overbalancing that of increasing temperatures. Similarly the internal energies of both liquid and solid increase along the melting line.

A large number of polymorphic transitions have also been studied in the new pressure range and in some cases a determination of all the thermodynamic parameters of the transition has been recorded. A statistical study of the results indicates a much greater tendency, in the new pressure range, for the occurrence of ‘ abnormal ’ transition lines—that is, lines with a negative *dt/dp* slope, which means that the high temperature phase has a smaller volume than the low temperature phase.

From the theoretical point of view it is disappointing to find that no consistent relations between constitution and compressibility have yet been noted even with hydrocarbon series and a wide range of organic isomers. That a relation does exist is almost certain but it may well be masked at the temperatures employed in the the present work.

Professor Bridgman's pioneerng work during the past 40 years has been so varied and extensive as to provide ample material to fill the present volume. It is inevitable, therefore, that little space could be devoted to the results of other workers in the field, and many significant contributions receive only brief mention. Thus, for example, the application of high pressures to large scale industrial processes has emphasized the need for more extensive pressure–volume–temperature data covering not only a wider range of pure substances but also multi-component systems in one or more phases. This work is now being undertaken in a number of laboratories and important papers have been published by A. Michels in Amsterdam, Sage and Lacey in California and Gibson in Washington. In addition considerable advances have been made to the theory of the liquid state and to the equation of state problem. Much of this work links up with that of Professor Bridgman and it is to be regretted that considerations of space have prevented a more detailed account of it being given.

D. M. N.

Faraday's Discovery of Electromagnetic Induction, by THOMAS MARTIN. Pp. 160. (London : Edward Arnold & Co., 1949.) 9*s*.

As is well known, Faraday kept a detailed diary covering his years of mature research at the Royal Institution, that is, from 1820 to 1862. This was first published at intervals between 1932 and 1936, edited by Thomas Martin from the manuscript in the possession of the Royal Institution. So voluminous a work—it is in seven volumes—is, however, accessible to comparatively few, so that all who are interested in Faraday's work and in the early history of electromagnetism will welcome the appearance of Mr. Martin's new book.

Faraday's published papers and the *Experimental Researches in Electricity* give in a masterly manner the results of his researches ; but the steps by which he was led to his epoch-making discoveries, the little revealing touches that show how his mind was working, these are to be found only in the Diary itself. Mr. Martin, by a judicious choice of extracts, has succeeded in conveying much of the charm and intimacy of the Diary. His book is not, however, merely an abridged version of the relevant portions of the larger work, but gives a connected account of the early history of electromagnetism from the announcement in 1820 of Oersted's discovery to the conclusion of Faraday's electromagnetic researches in the spring of 1832, the period covered by the first volume of the Diary in its printed form.

A brief survey of the main developments in electricity arising out of Galvani's discovery provides the historical setting for Faraday's work. This is followed by an account of Faraday's own early experiments in electricity—those on electromagnetic rotations, made in the latter part of 1821. After the successful conclusion of these experiments ten years elapsed before Faraday was again free to give his full attention to electricity, though scattered references in the Diary show how his thoughts dwelt on the subject. Then, between the end of August 1831 and the beginning of March 1832, came that amazing series of experiments and observations 'which may be said to constitute the discovery of electromagnetic induction'. Mr. Martin, in describing these researches, makes frequent use of Faraday's own words ; the diagrams are, with few exceptions, reproductions of Faraday's freehand sketches. At the same time much is to be found here that is not to be found in the Diary itself. For instance, we read (p. 59) that the art of covering copper wire with silk or cotton was apparently a monopoly in the hands of the opticians, who charged exorbitant prices ; that experimenters were advised to take their own copper wire to a bonnet-wire coverer living in Spitalfields, who worked at much lower rates ; iron wire could be obtained cotton-covered and was probably bonnet-wire used for stiffening ladies' bonnets.

We have nothing but praise for this modest little volume, and hope that Mr. Martin will deal in a similar manner with other aspects of Faraday's work. N. H. DE V. HEATHCOTE.

Giant Brains or Machines That Think, by E. C. Berkeley. Pp. xvi+270. 1st Edition. (New York : John Wiley & Sons, London : Chapman & Hall, 1949.) 32*s*.

This book is primarily concerned with large-scale calculating machines. It includes detailed information about IBM punched card machines (usually called Hollerith in this country after the originator), the Rockefeller differential analyser at M.I.T., the Automatic Sequence Controlled Calculator at Harvard University, the ENIAC built at the Moore School of Electrical Engineering, University of Pennsylvania, and machines using telephone relays built by the Bell Telephone Laboratories. With the exception of the IBM machines, which are widely used for commercial accounting, all these machines are the result of pioneer development work, and only one example of each (two in the case of the BTL machine) has been constructed.

In the first chapter the author is concerned with supporting the implication contained in the title that calculating machines can 'think'. Many people, however, will consider that he stretches the meaning of the word unduly. An automatic calculating machine can store numbers and perform a prearranged series of arithmetical operations on them. It is not, however, justifiable to say that such a machine can think, even though at certain stages the nature of the operation performed may depend in some way on the result of previous operations (e.g. on the sign of a number previously calculated) and may, therefore, be unknown to the operator. It is true that the sequences of operations performed by automatic calculating machines are much more complex in structure than the simple cycles performed by machines hitherto familiar—such as automatic lathes—and that they have suggested to those concerned lines of thought which may help to elucidate some aspects of animal behaviour. But the author, with his insistence on the statement that existing calculating machines can think, obscures rather than clarifies these important issues.

The book is intended for anyone who has a general interest in scientific matters. It calls for some readiness to follow involved arguments and descriptions. In places the treatment is somewhat condensed, and anyone unfamiliar with the subject would find consecutive reading rather exhausting ; the author, however, advises his readers not to do this, but to read what interests them, taking a bit here, and a bit there. Some sections would have been made easier to follow if technical terms had been introduced and defined only when they were essential for the later discussion. The style is informal but for the most part concise.

About one-third of the book is taken up with detailed descriptions of the machines mentioned above. Each machine is dealt with under the following headings : origin and development, general organization, physical devices used, input and output of information and the handling of it inside the machine. This is undoubtedly the most useful section of the book, and the author has brought together much information which was hitherto obtainable only from scattered papers. The book also contains some discussion of machines for performing the processes of symbolic logic, and some remarks on the social implications of the subject. A good feature is a bibliography of 20 pages.

The author has rightly dealt at greatest length with those machines which were completed and working at the time he was writing. It is, nevertheless, a pity that he did not say more about the EDVAC and machines like it. Although none of these machines were in operation until just before the book was published, the ideas on which they are based are original and even revolutionary ; moreover, a discussion of them would have brought out very clearly how, and within what limits, a machine can perform the highly complex sequences of operations which the author chooses to call 'thinking'.

M. V. WILKES.

Properties of Metals at Elevated Temperatures, by GEORGE V. SMITH. Pp. x+401. 1st Edition. (London and New York: McGraw-Hill, 1950.) 59*s*. 6*d*. ($7.00).

Almost all of this book is about the mechanical properties of metals at high temperatures, and especially about their creep properties. There is little about phase diagrams, diffusion, structure changes or conductivity. On the subject matter chosen, however, the book can be warmly recommended. About half of it deals with fundamental work on flow and fracture, both of single crystals and polycrystalline metals and alloys. There is a particularly interesting discussion of the concept of an equicohesive temperature and of the part played by grain boundaries in creep of metals. A short chapter on theories of flow and fracture which make use of models on the atomic scale such as that of dislocations is scarcely adequate to the progess made in this subject in the last few years ; but research workers in this field will find summarized in this book just the type of experimental material on which successful theories should be based.

The second half of the book deals with the technical side of this extensive and complicated subject; there are chapters on tests and test equipment, effect of chemical composition, heat treatment and so on, and on design for elevated temperature service.

N.F.M.

Acoustic Measurements, by L. L. BERANEK. Pp. vi+914. 1st Edition. (New York: John Wiley & Sons; London : Chapman and Hall, 1949.) 56*s*.

A laboratory manual of acoustics, which will give the research worker details of all the latest techniques has been a long-felt need and Dr. Beranek has ably filled the gap with this bulky handbook which runs to 900 pages. He has drawn on techniques developed in part by himself at M.I.T. in special war-time assignments as well as on the work of well known firms such as the Bell Telephone Co. Although a considerable amount of this material is to be found in technical journals, it has not been grouped in this useful form before and some of it, indeed, has never been published.

The subjects cover most of the measurements in air that one would be likely to require in an audio-frequency laboratory. Starting with a short theoretical chapter on propagation and a little physiology of the ear, the author ranges over impedance measurement, sound analysers, tests of microphones, loudspeakers and communication systems to auditorium acoustics and the measurement of acoustical materials.

With so many aspects of technical acoustics mentioned, it is perhaps churlish to ask for more, but we should like in a second edition to read something of the special techniques required in underwater measurements—the hydrophone is barely mentioned—and open-air acoustics.

Dr. Beranek is perhaps too firmly wedded to the electrodynamic or electrostatic transducer as measuring instrument, as opposed to those used in fluid dynamics of which he seems to have less experience. Thus, in a comprehensive treatment of the methods of measuring acoustic impedance he rejects the direct measurement of particle velocity at a point as being too difficult though, in our experience, this factor is easier to measure than pressure amplitude in a confined space, owing to the obstruction offered by a microphone. On the other hand, the measurement of the direct flow of air through an absorbent specimen is described as a simple test to give information about the acoustic properties of the specimen. Even if the D.C. and A.C. hydrodynamics of a porous material were comparable—and this matter is not discussed—we doubt whether many textile workers or rheologists would admit that the results of such tests were easy of interpretation.

Generally, in fact, the practical man would like a more critical approach to the various techniques described. The author, perhaps through modesty, hesitates to say which of the various ways of making an acoustic test he favours. He should give the reader the benefit of his considerable experience in this respect. E. G. R.

Physical Properties of Glass, by J. E. STANWORTH. Pp. vii + 224. Monographs on the Physics and Chemistry of Materials. (Oxford: University Press, 1950.) 21*s*.

As a general survey of what has been published during the past 20 to 30 years about the physical properties of glass, this book serves a useful purpose. The survey is wide though not exhaustive, and it is put together in an easily readable manner. The matter presented is open to much criticism, but where this is offered it must be taken as criticism of the views of the original authors. The general theme followed by the author in selecting the material dealt with is the correlation of the physical properties with the fundamental structure of glass.

In certain sections the published work has consisted of little more than accounts of the results obtained in experimental investigations, sometimes expressed by empirical relationships or illustrated by mathematical investigations of hypothetical mechanical systems which contribute little to the better understanding of the fundamental nature of glass or glass structure. Many of the experimental results are well authenticated and are sufficient in themselves to indicate the complex character of glass. This indication is strengthened by the marked divergences which are not infrequently found between the results obtained by equally competent experimental workers.

Where the authors have attempted to deduce fundamental information from their results the reliability of their deductions varies from 'highly probable' to 'highly speculative'. This range of reliability is due to several factors, not the least important of which is that in many cases generalized conclusions, with theories to suit, have been drawn from observations made on small numbers of glasses covering very limited ranges of compositions, and sometimes consisting only of binary glasses of the alkali–silica or alkali–boric oxide types. In formulating their theoretical views, also, the authors have not in all cases considered it necessary to be bound by the views which are generally accepted as well-established in 'ordinary' chemistry, nor have they always considered how far their views are consistent with properties of glass other than those which they have themselves been investigating. Fortunately, this is not universally true, and in a proportion of the work quoted, not only the experimental results but also the suggestions and conclusions arising out of them must be accepted with due respect. The survey makes it abundantly clear, however, that really reliable knowledge of the fundamental nature of glass is very scanty, and that speculation has in many directions gone far beyond the stage warranted by the experimental evidence available.

The author draws attention in many places to the need for additional information on specific properties and, perhaps wisely, does not attempt to assess the relative merits of experimental results where these are not in agreement, nor, for the most part, does he express opinions for or against the theoretical views put forward. H. MOORE.

CONTENTS FOR SECTION A

PAGE

Dr. A. S. Russell. Lord Rutherford, 1907–19 : a Partial Portrait 217

Dr. W. J. Swiatecki. The Nuclear Surface Energy 226

Dr. J. W. Gardner. Directional Correlation between Successive Internal-Conversion Electrons : II 238

Mr. R. Taylor. Complete Molecular Orbital Treatment of the System H_4 . . 249

Dr. R. McWeeny. The Diamagnetic Anisotropy of Large Aromatic Systems : Parts I and II 261

Dr. G. C. Benson and Dr. G. Wyllie. A Quantum Mechanical Treatment of the Lithium Fluoride Crystal 276

Prof. L. Brú and Dr. M. K. Gharpurey. Epitaxial Crystal Growth of Silver on Rock-Salt (110) and (111) Faces 283

Dr. S. Levine and Mr. A. Suddaby. Simplified Forms for Free Energy of the Double Layers of Two Plates in a Symmetrical Electrolyte 287

Dr. A. Papapetrou. Equations of Motion in General Relativity : II—The Coordinate Condition 302

Letters to the Editor :

- Dr. A. G. Gaydon and Dr. H. G. Wolfhard. The Excitation of HCO, C_2 and CH Bands in Flame Spectra 310
- Mr. W. Mims and H. Halban. Positron Electron Branching 311
- Mr. F. K. Goward and Mr. J. J. Wilkins. Detection of $^{14}N(\gamma, d\ 3\alpha)$ and $^{14}N(\gamma, pn\ 3\alpha)$ Reactions in Nuclear Emulsions 312
- Dr. Joan M. Freeman and Mr. J. Seed. Alpha-Particles from the Proton Bombardment of Phosphorus 313
- Dr. B. Bleaney. On the Spatial Alignment of Nuclei 315
- Dr. B. Bleaney. Note on the Adiabatic Demagnetization of Cobalt Ammonium Sulphate 316
- Prof. E. A. Stewardson and Mr. P. A. Lee. The M_{IV} and M_V Absorption Edges of ^{68}Er 318

Reviews of Books 319

Contents for Section B 325

Abstracts for Section B 325

ABSTRACTS FOR SECTION A

5th Rutherford Lecture, by A. S. Russell.

The Nuclear Surface Energy, by W. J. Swiatecki.

ABSTRACT. The kinetic and potential parts of the nuclear surface energy are calculated on the basis of the statistical model of the nucleus taken to first order and using a gaussian interaction between particles. The resulting surface energy is found to be of the right order of magnitude, the calculated figure being in fact 55% of the empirical value.

Directional Correlation Between Successive Internal-Conversion Electrons: II, by J. W. Gardner.

ABSTRACT. A previous paper examined the directional correlation $I(\theta)$ between two conversion electrons in cascade, each electron being ejected, with energy much less than mc^2, from an initial s-state, under the action of a pure electric multipole γ-transition, and $I(\theta)$ was shown to be completely determined by symmetry arguments alone in this case. In the present paper we consider the same problem for pure magnetic, and for mixed transitions. Here $I(\theta)$ usually appears as a weighted mixture of constituents $I_1(\theta)$, $I_2(\theta)$, etc., each of which is completely determined by symmetry arguments alone. The weight factors, however, must be determined from standard internal-conversion theory.

Complete Molecular Orbital Treatment of the System H_4, by R. Taylor.

ABSTRACT. A complete molecular orbital calculation has been made for the linear system of four protons with four associated electrons. Results indicate the validity of the assumptions usually made in approximate calculations of this type. All the three- and four-centred integrals usually neglected in such work have been evaluated and found to be of great importance in determining the energy levels of the system.

The Diamagnetic Anisotropy of Large Aromatic Systems: Parts I and II, by R. McWeeny.

ABSTRACT.

Part I. A short discussion is given of the magnetic properties of large aromatic hydrocarbons and of the applicability of various wave-mechanical methods of calculating the diamagnetic susceptibility. The ultimate particles of most cokes and carbon blacks are not sufficiently large to be regarded as crystalline and must be treated as large molecules: calculations on certain simple types of molecule are to be carried out with a view to elucidating the effect of particle size and shape on the magnetic properties of such carbons. For systems of these dimensions the molecular orbital method should be valid and, in fact, appears to constitute the only feasible approach to the problem.

Part II. The method of molecular orbitals is applied in a discussion of the diamagnetic anisotropy of the p-polyphenyls, which may be regarded as the simplest type of aromatic chain molecule. An almost perfectly linear relationship is predicted between diamagnetic anisotropy and chain length: in this case the molecular orbital method should be valid for chains of up to at least 2,000 rings. The results are discussed and the desirability of further accurate experimental work on these molecules is pointed out: this would make possible a stringent test of the whole theory.

A Quantum Mechanical Treatment of the Lithium Fluoride Crystal, by G. C. Benson and G. Wyllie.

ABSTRACT. A simplified quantum-mechanical model for the LiF crystal is described. The lattice parameter, cohesive energy and compressibility are calculated using three different choices of wave function for the fluorine ion in the crystal; these are (i) the Hartree

wave functions for the free ion F^-, (ii) wave functions orthogonalized to the 1s-orbitals of the six nearest neighbour lithium ions, (iii) functions similar to (ii) but based on wave functions for the free ion which have been contracted to give agreement with the diamagnetism of F^-.

The results obtained in (i) are in fair agreement with the empirical values. Large discrepancies are observed when the more complete wave functions (ii) are employed. This is taken as indicating that the Hartree wave functions for the free ion are too diffuse and the results of the third calculation, which agree well with experiment, substantiate this interpretation.

Epitaxial Crystal Growth of Silver on Rock-Salt (110) *and* (111) *Faces*, by L. BRÚ and M. K. GHARPUREY.

ABSTRACT. It has long been known that silver can grow epitaxially on a rock-salt cleavage face. It has now been found by electron diffraction that silver also grows in parallel orientation on (110) and (111) faces of rock-salt and that the silver crystals are twinned repeatedly on their {111} faces. A new example of rotational slip has been noted in which silver crystals were found to have slipped rotationally on their {110} planes.

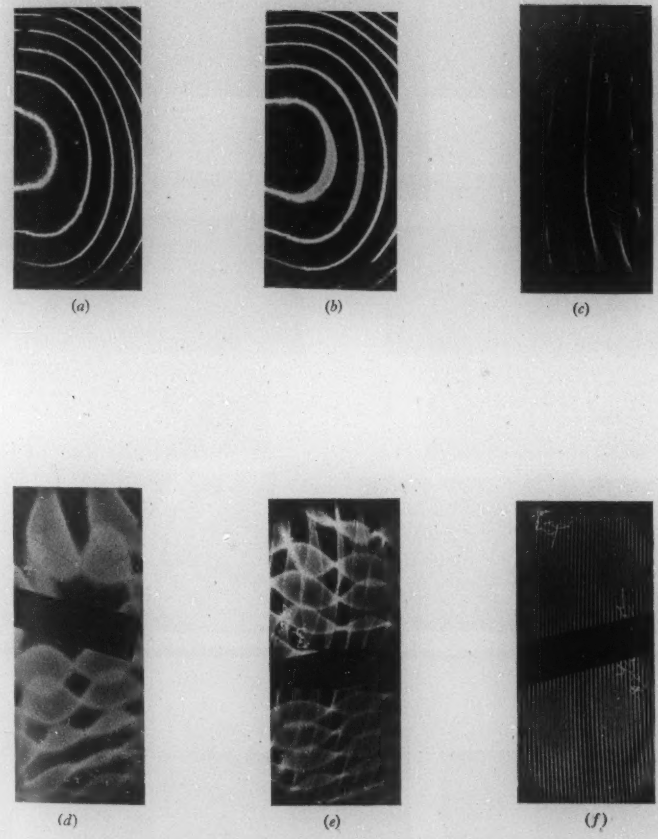
Simplified Forms for Free Energy of the Double Layers of Two Plates in a Symmetrical Electrolyte, by S. LEVINE and A. SUDDABY.

ABSTRACT. Approximate formulae are obtained for the interaction (free) energy of the electric double layers of two parallel plates immersed in an electrolyte of the binary, symmetrical valency type. The three separate simplified expressions which are derived are good approximations at (i) large separations between the plates, (ii) small separations between the plates, and (iii) moderate potentials. The first form is an extension of a result proposed by Verwey and Overbeek in their recent book on the theory of the stability of lyophobic colloids. The third form is a power series in the surface potential, the first term of which is identical with the expression obtained when the linear Debye–Huckel equation in the theory of electrolytes is used. These results are more suitable for numerical computations than the exact formula for the free energy, which is expressed in terms of elliptic integrals.

Equations of Motion in General Relativity: II—The Coordinate Condition, by A. PAPAPETROU.

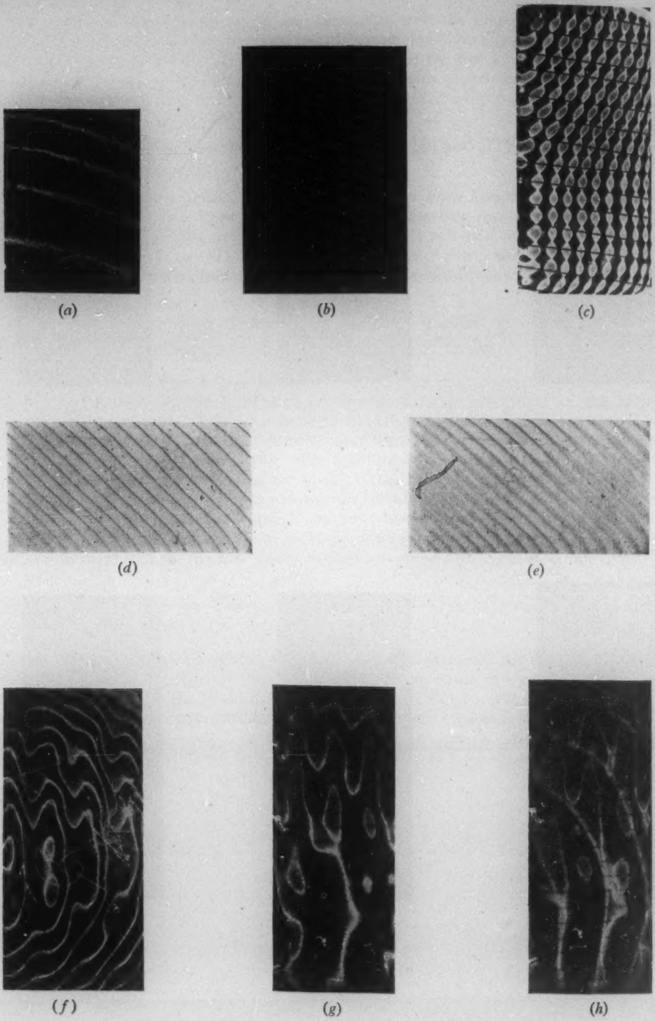
ABSTRACT. The paper contains a discussion of the necessity for using a coordinate condition in the calculations for the derivation of the equations of motion from the field equations. It is found that no coordinate condition is necessary in the first approximation; i.e. one gets the Newtonian equations of motion if one merely assumes that the gravitational field is weak and quasi-static. In the higher approximations a coordinate condition is indispensable in order to eliminate all apparent accelerations from the equations of motion.

PROC. PHYS. SOC. VOL. 64, PT. 3—B (S. TOLANSKY AND W. BARDSLEY)


(a) (b) (c)

(d) (e) (f)

PLATE I.

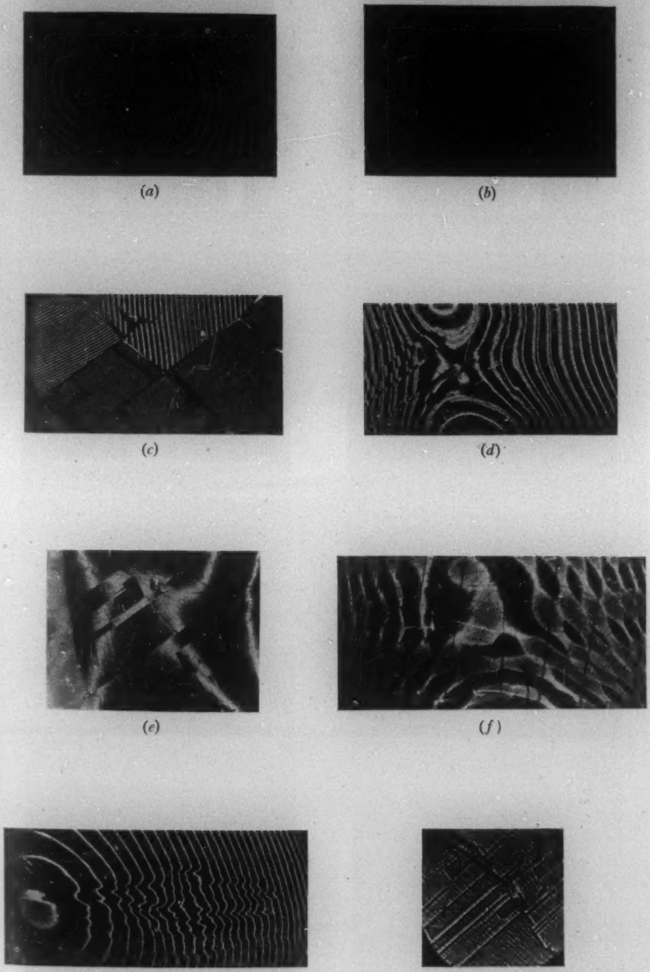


(a) (b) (c)

(d) (e)

(f) (g) (h)

PLATE II.



(a) (b)

(c) (d)

(e) (f)

(g) (h)

PLATE III.

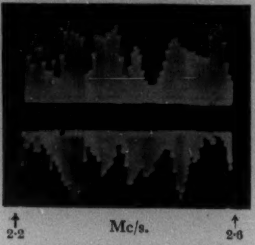


Figure 6.

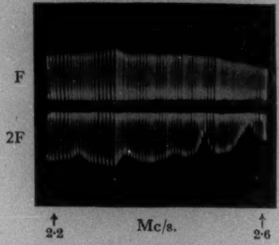


Figure 7.

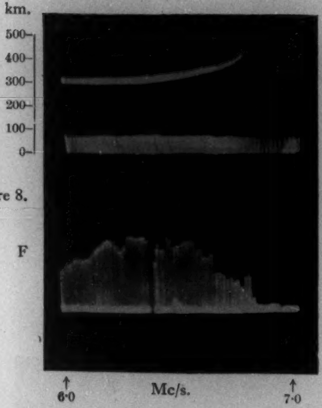


Figure 8.

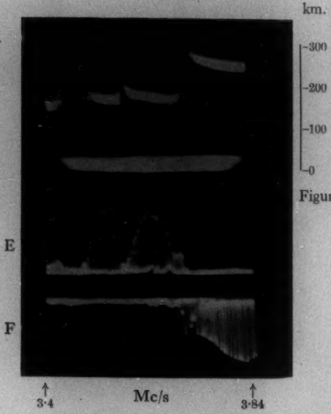


Figure 9.

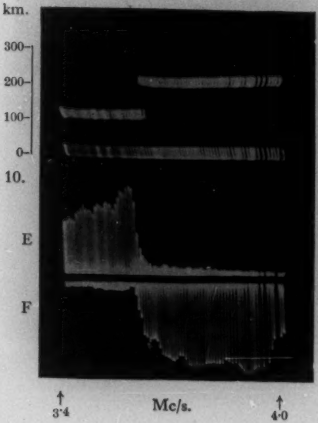


Figure 10.

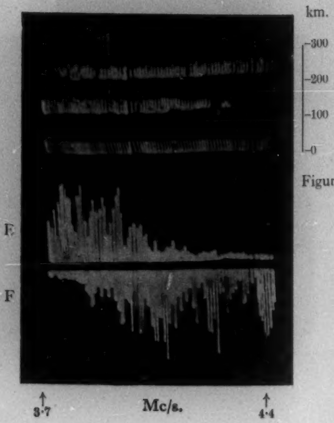
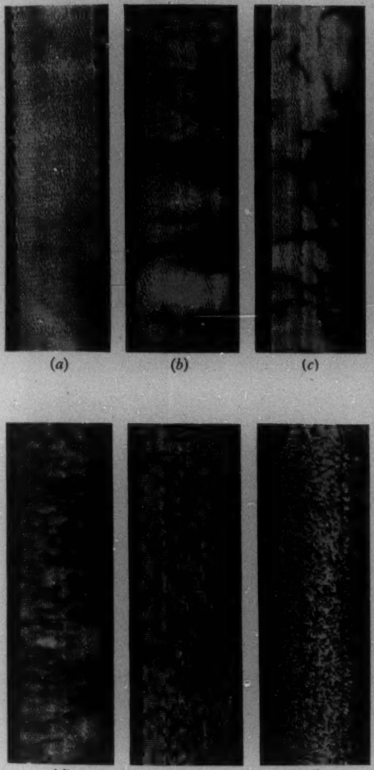


Figure 11.

PLATE I.


(a) (b) (c)

(d) (e) (f)

Figure 13.

Short sections of panoramic records for region E obtained for the same frequency range of 2·2 to 2·6 Mc/s. on different occasions. Each occupies four minutes.

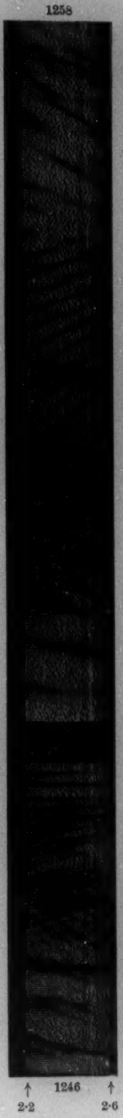


Figure 14.
25.11.49

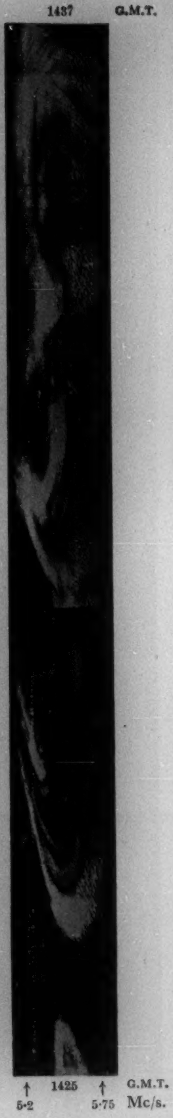


Figure 15.
24.5.49

PLATE II.

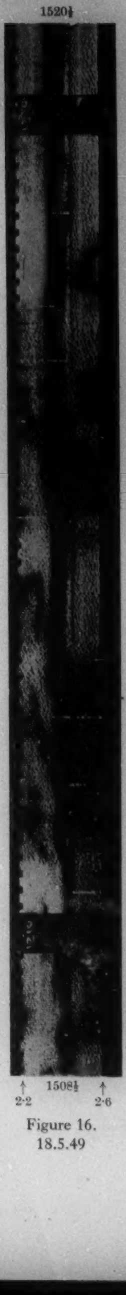

1520½
↑ 2·2
1508½
↑ 2·6

Figure 16.
18.5.49

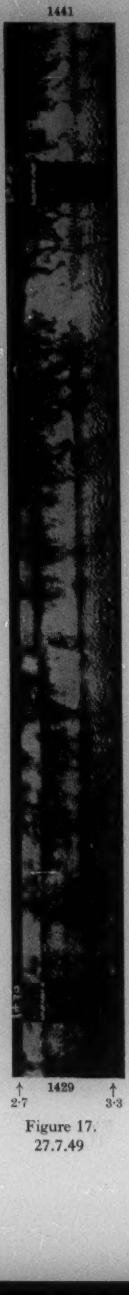

1441
↑ 2·7
1429
↑ 3·3

Figure 17.
27.7.49

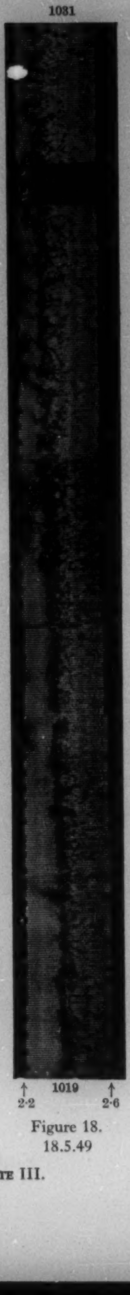

1031
↑ 2·2
1019
↑ 2·6

Figure 18.
18.5.49

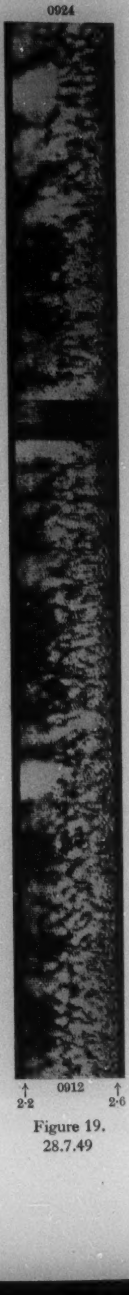

0924
G.M.T.
↑ 2·2
0912
↑ 2·6
G.M.T.
Mc/s.

Figure 19.
28.7.49

PLATE III.

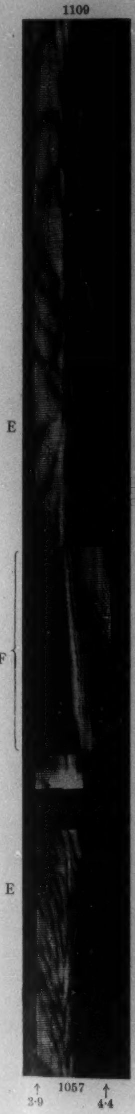


Figure 20.
Ab. E and F
24.5.50

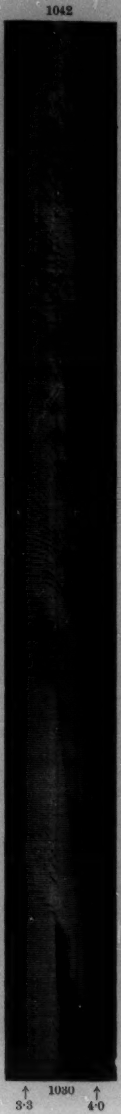


Figure 21.
Ab. E
31.7.50

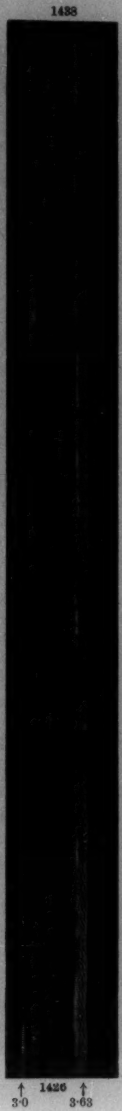


Figure 22.
E and Ab. E
2.6.50

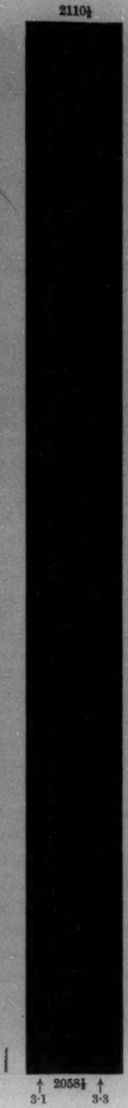


Figure 23.
F
31.5.50

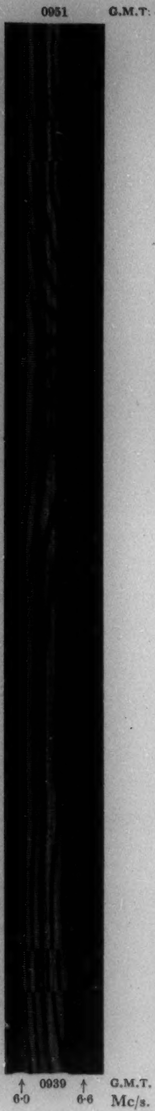


Figure 24.
Ab.E
5.6.50

PLATE IV.